# ROBUST MODEL-FREE VARIABLE SCREENING, DOUBLE-PARALLEL MONTE CARLO AND AVERAGE BAYESIAN INFORMATION CRITERION

A Dissertation

by

JINGNAN XUE

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

| | |
|---|---|
| Chair of Committee, | Samiran Sinha |
| Committee Members, | Bani Mallick |
| | Anirban Bhattacharya |
| | Jianxin Zhou |
| Head of Department, | Valen Johnson |

August 2017

Major Subject: Statistics

# ABSTRACT

Big data analysis and high dimensional data analysis are two popular and challenging topics in current statistical research. They bring us a lot of opportunities as well as many challenges. For big data, traditional methods are generally not efficient enough to handle them, from both time perspective and space perspective. For high dimensional data, most traditional methods can't be implemented, let alone maintain their desirable properties, such as consistency.

In this disseration, three new strategies are proposed to solve these issues. HZ-SIS is a robust model-free variable screening method and possesses sure screening property under the ultrahigh-dimensional setting. It works based on the nonparanormal transformation and Henze-Zirkler's test. The numerical results indicate that, compared to the existing methods, the proposed method is more robust to the data generated from heavy-tailed distributions and/or complex models with interaction variables.

Double Parallel Monte Carlo is a simple, practical and efficient MCMC algorithm for Bayesian analysis of big data. The proposed algorithm suggests to divide the big dataset into some smaller subsets and provides a simple method to aggregate the subset posteriors to approximate the full data posterior. To further speed up computation, the proposed algorithm employs the population stochastic approximation Monte Carlo (Pop-SAMC) algorithm, a parallel MCMC algorithm, to simulate from each subset posterior. Since the proposed algorithm consists of two levels of parallel, data parallel and simulation parallel, it is coined as "Double Parallel Monte Carlo". The validity of the proposed algorithm is justified both mathematically and numerically.

Average Bayesian Information Criterion (ABIC) and its high-dimensional variant Average Extended Bayesian Information Criterion (AEBIC) led to an innovative way to use posterior samples to conduct model selection. The consistency of this method is established for the high-dimensional generalized linear model under some sparsity and regularity conditions. The numerical results also indicate that, when the sample size is large enough, this method can accurately select the smallest true model with high probability.

# ACKNOWLEDGMENTS

First and foremost I would like to express my sincerest gratitude to my advisor, Dr. Faming Liang for his excellent guidance, systematic mentoring, and constant support throughout my entire doctoral study. I'm very lucky to have him as my advisor. Without his help, I would not be able to finish this dissertation. His continuous passion in research and deep understanding of statistical methodology also let me know what properties one should possess to achieve success in academia.

My gratitude also goes to my committee chair, Dr. Samiran Sinha, for assisting me to handle documents with patience, and other committee members, Dr. Bani Mallick, Dr. Anirban Battacharya, Dr. Jianxin Zhou for generously giving insightful comments and suggestions to improve my dissertation work. I'm very honored to have them on my Ph.D. advisory committee.

I would also like to express my thanks to Dr. Michael Longnecker for his constant help ever since I entered the department. As the associate department head, he is extremely caring and nice to every graduate student. I'm also thankful to Dr. Jianhua Huang, Dr. Jeffery Hart, Dr. Willa Chen, Dr. Ursula Mueller, Dr. Valen Johnson, Dr. Mohsen Pourahmadi, Dr. Huiyan Sang, Dr. Suhasini Subba Rao, Dr Edward Jones and Dr. Raymond Carroll. I really learned a lot from their courses.

Last but not least, I would like to thank my parents for providing me unconditional support and all my friends for making my past five years colorful and memorable. I'll cherish this for the rest of my life.

# CONTRIBUTORS AND FUNDING SOURCES

## Contributors

This work was supervised by Professor Faming Liang and a dissertation committee consisting of Professor Samiran Sinha (chair), Professor Bani Mallick, Professor Anirban Bhattacharya from the Department of Statistics and Professor Jianxin Zhou from the Department of Mathematics.

All work conducted for the dissertation was completed by the student independently.

## Funding Sources

Graduate study was supported by a fellowship from Texas A&M University and in part by the NSF grants DMS-1545202 and DMS/NIGMS R01-GM117597.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

# 1. INTRODUCTION

Big data and High dimensional data are two popular and challenging topics in current statisticial research. Both have attracted lots of attention and lead to numerous innovative methods. This dissertation contributes three new approaches, the first and the third one can be used to deal with high dimensional data, and the second one is developed for big data analysis.

## 1.1 High Dimensional Data

In statistics, high-dimensional data usually refers to the data whose dimension p is larger than the sample size n and ultrahigh-dimensional data usually refers to the data where $p$ grows exponentially with $n$, that is, $p = O(n^{\alpha})$.

In recent years, high dimensional data analysis has become more frequent and important in diverse fields of our daily life, such as genomics, microarrays, proteomics, brain images, climatology, geology, neurology, health science,economics, finance and machine learning (Fan and Lv, 2010). For example, in genowide association studies between genotypes and phenotypes, millions of SNPs are potential covariates; in disease classfication using microarray or proteomics data, thousands of expression profiles are potential predictors; in biomedical and clinical studies, a large number of magnetic resonance images (MRI) and functional MRI data are collected for each subject. Moreover, when interaction are considered, the dimensionality will grow more quickly. These massive amounts of high dimensional data have undoubtedly brought many opportunities for scientific development. However, at the same time, they have also significantly challenged traditional statistical theory (Fan and Li, 2007; Johnstone and Titterington, 2009).

One big challenge comes from the collinearity of the predictors. As we all know,

even in the low dimensional case, the notorious collinearity will sometimes cause lots of trouble, and force us to use PCA or any other techniques to solve it. In the high dimensional setting, this problem becomes worser and usually makes us to overfit the data and select wrong models, since any variable can be well approximated or even replaced by a combination of spurious variables. Besides, under the high dimensional setting, some traditional methods can't even be implemented. For example, in gaussian graphical models (GGM), the sample covariance matrix is singular when $p > n$, thus we can no longer directly estimate the concentration matrix and use it to learn the graph. (Liang *et al.*, 2015) Noise accumulation in high dimensional prediction has been recongnized as another challenge. In fact, in supervised learning problems, prediction using all features can be as bad as random guess. Therefore, variable selection is fundamentally important to high dimensional data analysis, for both regression and classification. To overcome these challenges and make high dimensional statistical inference possible, we usually impose the so called sparsity condition. For example, in supervised learning, that means most of the features are irrelevant with the response variable; in Gaussian Graphical Models, that means only few edges truly exist. With sparsity assumption and specific methods, variable selection is able to possess the consistency property and can be implemented to improve the model interpretability and prediction accuracy.

In the past decades, numerous innovative methods have been proposed to deal with high dimensional data. The frequentist methods are usually regularition-based, imposing a penalty on the likelihood function to enforce sparsity. For example, Tibshirani (1996) employs a $l_1$-penalty, Zou and Hastie (2005) employs a combination of $l_1$ and $l_2$ penalties, Fan and Li (2001) employs a smoothly clipped absolute deviation penalty, Zhang (2010) employes minimax concave penalty and Song and Liang (2015) employes a reverse $l_1$ penalty. The Bayesian methods usually employ appro-

priate prior distributions to enforce sparsity, such as the non-local priors (Johnson and Rossell, 2012), the priors used in EBIC (Liang *et al.*, 2013; Chen and Chen, 2008, 2012), etc.

In this disseration, we contribute two more methods. In Chapter 2, we'll introduce a robust model-free variable screening method. It is well known that variable screening can reduce the dimension of feature space to a moderate scale while keeping all relevant features. Thereby it can act as a preliminary step ahead of variable selection. Compared to the previous proposed model-free variable screening methods, which more or less require some additional assumptions and are not robust to non-regular data, our method only needs few assumptions to guarantee its sure screening property under ultrahigh-dimensional setting, and is more robust to the heavy-tailed data and data with interaction effects. In Chapter 4, we'll describe a new information criterion, ABIC and its high dimensional variant AEBIC. To our knowledge, AEBIC is the first information criterion that use information from posterior samples and possess model selection consistency.

## 1.2 Big Data

Big data often refers to the data whose the sample size is too large such that it can not be processed and stored on a single computer.

Nowadays, we are entering the era of Big data, thanks to the technology development and information explosion. Big data can be found almost everywhere around our life. For example, Walmart handles more than 1 million customer transactions every hour, which are imported into databases estimated to contain more than 2.5 petabytes of data - the equivalent of 167 times the information contained in all the books in the US Library of Congress; Amazon.com handles millions of back-end operations every day, as well as queries from more than half a million third-party

sellers; Windermere Real Estate uses anonymous GPS signals from nearly 100 million drivers to help new home buyers determine their typical drive times to and from work throughout various times of the day.

Big data brings opportunities to modern society. It helps us discover many new findings which are not possible with small-scale data. At the same time, the massive sample size also brings unique challenges to computational and statistical research. (Fan *et al.*, 2014) They mainly come from two aspects. One is on data storage and the other is on computational time. Although hardware is being upgraded in an unprecedented speed to enlarge storage memory and decrease computational time, more efficient algorithms are still highly required to solve these challenges.

Thanks to the advent of parallel computing, a large number of efficient approaches have now been proposed to deal with Big data. For example, it is well known that when making Bayesian inference for big data, traditional MCMC algorithms are generally not efficient enough. Therefore, people propose to divide the large dataset into a number of smaller subsets, and then conduct the Bayesian analysis for each subset separately. Finally, the posterior samples generated for each subset are aggregated in some way such that a correct inference can be made for the full data posterior. Several algorithms have been developed to address the issue of subset posterior aggregation, such as Scott *et al.* (2016); Neiswanger *et al.* (2013); Wang and Dunson (2013); Minsker *et al.* (2014); **?**); Srivastava *et al.* (2015)

In Chapter 3, we introduce a new method for aggregating subset posterior samples. The new method is surprisingly simple, which is to first simulate from some modified subset posteriors, for which the log-likelihood functions are appropriately scaled according to their sample size, and then recenter the subset posterior samples to their global mean. In order to further speed up computation, we suggest to use the Pop-SAMC algorithm (Song *et al.*, 2014), rather than traditional single chain

MCMC algorithms, to draw samples from each subset posterior. Since the proposed method consists of two levels of parallel, data parallel and simulation parallel, it is coined as "double parallel" Monte Carlo.

### 1.3 Dissertation Structure

The rest of the dissertation is organized as follows. Chapter 2 introduces a robust model-free variable screening method to deal with ultra-high dimensional data. Chapter 3 develops the Double-Parallel Monte Carlo algorithm for Bayesian analyisis of big data. Chapter 4 is dedicated to the Average Bayesian Information Criterion(ABIC) and its high-dimensional variant Average Extended Bayesian Information Criterion(AEBIC). Chapter 5 gives a summary of this dissertation and points out some directions for future research. The technical details and supplementary results are included in the Appendix.

# 2. ROBUST MODEL-FREE FEATURE SCREENING FOR ULTRAHIGH DIMENSIONAL DATA

## 2.1 Introduction

Variable selection plays an important role in high-dimensional data analysis. However, under the ultrahigh-dimensional setting, where the number of covariates may grow at an exponential rate of sample size, current variable selection methods may not work well due to the simultaneous challenges of computational expediency, statistical accuracy, and algorithmic stability (Fan *et al.*, 2009). A practical approach is using a screening procedure to reduce the dimension of feature space to a moderate scale, and then implementing variable selection methods on the reduced dataset. To pursue this approach, Fan and Lv (2008) proposed the sure independence screening (SIS) method for linear regression, which screens predictors by ranking their Pearson correlations with the response variable. They established the sure screening property, that is, all active predictors can be selected with probability approaching one as the sample size increases to infinity. Fan and Song (2010) extended SIS to generalized linear models, which screens features by ranking the maximum marginal likelihood estimators. As for nonlinear feature screening, Hall and Miller (2009) suggested polynomial transformations of predictors, and Fan *et al.* (2011) suggested to estimate the nonparametric components marginally using B-splines and then screen the features by ranking the magnitude of nonparametric components.

All the above methods require the specification of a particular model structure. If the underlying model is correctly specified, these methods can perform well. However, if the underlying model is misspecified, their performance may be corrupted. Under the ultrahigh dimensional setting, specifying a correct model is usually an

impossible task, thus the model-free feature screening methods are often appealing. Toward this direction, Zhu *et al.* (2011) proposed a sure independence ranking and screening (SIRS) method to screen significant features for multi-index models. Li *et al.* (2012) proposed a distance correlation sure independence screening (DC-SIS) method, which screens features based on their distance correlations (Székely *et al.*, 2007) with the response variable. It is known that for two random variables, zero distance correlation implies independence. He *et al.* (2013) proposed a quantile-adaptive sure independence screening (Qa-SIS) method, which employs spline approximation to model the marginal effects of the predictors at a quantile level and then uses it to screen variables. A particular strength of this method is that it can handle the censored data arising in survival analysis. Recently, Cui *et al.* (2015) established a mean variance sure independence screening (MV-SIS) method, where the dependence of two random variables is measured using the mean variance of the conditional distribution function. This method is originally proposed for categorical response variables, but can be extended to the problems for which the response variable is continuous via discretization.

Although the model-free variable screening methods avoid the specification of a particular model structure, they are still based on some assumptions for the predictor and response variables, more or less. For example, DC-SIS requires both the predictors and the response variable to satisfy the sub-exponential tail probability uniformly. That is, practically, the response variable and predictors should be uniformly bounded or follow a multivariate Gaussian distribution. Qa-SIS requires the conditional quantile function's derivative to satisfy a Lipschitz condition and the conditional density function to be uniformly bounded for each feature.

In this chapter, we introduce a new model-free feature screening method and establish its sure independence screening property under the ultrahigh dimensional

setting. The proposed method works based on the nonparanormal transformation (Liu *et al.*, 2009) and Henze-Zirkler's test (Henze and Zirkler, 1990). It is to first transform the response variable and each of the predictors to Gaussian random variables using the nonparanormal transformation and then test the dependence between the response variable and the predictors using the Henze-Zirkler's test. Compared to the existing methods, the proposed method requires fewer assumptions to guarantee its sure independence screening property and thus performs more robustly. Our numerical studies indicate that the new method can achieve better performance when the covariates follow a heavy-tailed distribution and when the underlying true model is complex with interaction variables.

The rest of this chapter is organized as follows. In Section 2.2, we describe the proposed method and establish its sure independence screening property. In Section 2.3, we conduct simulation studies to evaluate the finite sample performance of the proposed method along with comparisons with the existing methods. In Section 2.4, we apply the proposed method to screening of anticancer drug response related genes.

## 2.2 Robust Feature Screening

### 2.2.1 The Method

Let $Y$ denote the continuous response variable, let $\boldsymbol{X} = (X_1, \ldots, X_p)$ denote the continuous covariates, let $n$ denote the sample size of the data, and let $f(y|\boldsymbol{x})$ denote the conditional distribution of $Y$ given $\boldsymbol{X}$. Under the ultrahigh-dimensional setting, where $p = O(\exp(n^\tau))$ for some $\tau > 0$, we generally assume that only few predictors are relevant to the response variable, although the covariate dimension $p$ greatly exceeds the sample size $n$. Without specifying a parametric form for the regression

model, we define the sets of active predictors and inactive predictors as follows:

$$\begin{aligned} D &= \{k : f(y|\boldsymbol{x}) \text{ functionally depends on } X_k\}, \\ I &= \{k : f(y|\boldsymbol{x}) \text{ does not functionally depends on } X_k\}. \end{aligned}$$

Directly identifying the active predictor set $D$ is sometimes difficult or even impossible under the ultrahigh-dimensional setting. Therefore, people proposed to first find a larger set with a moderate size including all elements in $D$, then apply variable selection techniques on this larger set to accurately identify $D$. Note that if $f(y|\boldsymbol{x})$ functionally depends on $X_k$, then $Y$ and $X_k$ are usually marginally dependent as well, therefore we can select the marginally dependent predictors to construct the larger set, which is usually referred to as independence screening. In fact, under the partial orthogonality condition (Fan and Song, 2010; Huang *et al.*, 2008), that is, $\{x_i, i \in D\}$ is independent of $\{x_j, j \in I\}$, we can further show that $f(y|\boldsymbol{x})$ functionally depends on $x_k$ if and only if $Y$ and $X_k$ are marginally dependent.

To implement independence screening, we need to find a metric to measure the marginal dependence between each predictor $X_k$ and the response variable $Y$. Several metrics have already been proposed, see e.g., Zhu *et al.* (2011), Cui *et al.* (2015), Li *et al.* (2012), and He *et al.* (2013). In this chapter, we propose a new one with the basic idea described as follows. Let $F_y(\cdot)$ denote the CDF of the response variable $Y$, and $F_k(\cdot)$ denote the CDF of the $k$th predictor $X_k$. Consider the nonparanormal transformation (Liu *et al.*, 2009)

$$T_y(Y) = \Phi^{-1}(F_y(Y)), \quad T_k(X_k) = \Phi^{-1}(F_k(X_k)), \quad k = 1, \ldots, p, \tag{2.1}$$

where $\Phi(\cdot)$ denotes the CDF of the standard Gaussian distribution. Notice the

nonparanormal transformation implemented here is slightly different with the original one because we don't need the mean and variance correction step. In addition, Liu *et al.* (2009) imposed nonparanormal transformation on nonparanormal distributions, but here we just use this transformation on common random vectors.

It is easy to see that $Y$ is independent of $X_k$ if and only if $(T_y(Y), T_k(X_k))$ forms a bivariate random vector following the distribution $N_2(\mathbf{0}, \mathbf{I_2})$, where $\mathbf{I_2}$ denotes the $2 \times 2$ identity matrix. The latter can be tested using a multivariate normality test, e.g., Henze-Zirkler's test (Henze and Zirkler, 1990), with the known covariance structure. If $(T_y(Y), T_k(X_k))$ does not follow the distribution $N_2(\mathbf{0}, \mathbf{I_2})$, then the Henze-Zirkler's test statistic tends to take a large value. In practice, since $F_y$ and $F_k$'s are usually unknown, we can use the estimated nonparanormal transformation in Liu *et al.* (2009). The estimated nonparanormal transformation has been implemented in the R package *huge*.

In summary, the proposed method consists of the following steps:

1. Transform all variables, including the response variable and predictors, to standard Gaussian random variables by the estimated nonparanormal transformation. Let's take the response variable as example, let

$$\widetilde{T}_y(y_i) = \Phi^{-1}(\widetilde{F}_y(y_i)), \quad i = 1, \ldots, n,$$

where $y_i$ denotes the $i$th observation of $Y$, $\widetilde{F}_y$ is the truncated empirical distribution of $Y$ given by

$$\widetilde{F}_y(t) = \begin{cases} \delta_n & : \widehat{F}_y(t) < \delta_n, \\ \widehat{F}_y(t) & : \delta_n \leq \widehat{F}_y(t) \leq 1 - \delta_n, \\ 1 - \delta_n & : \widehat{F}_y(t) > 1 - \delta_n, \end{cases}$$

$\widehat{F}_y(t) \equiv \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{y_i \le t\}}$ is the empirical distribution of $Y$, and the default truncation parameter $\delta_n = \frac{1}{4}n^{-1/4}(\pi \log n)^{-1/2}$.

2. For each predictor $X_k$, calculate the Henze-Zirkler test statistic

$$\widetilde{\omega}_k^* = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} e^{-\frac{\beta^2}{2}d_{ij}} - \frac{2}{n(1+\beta^2)}\sum_{i=1}^{n} e^{-\frac{\beta^2}{2(1+\beta^2)}d_i} + \frac{1}{1+2\beta^2}, \tag{2.2}$$

where $\beta$ is the smoothing parameter and its optimal value is $(1.25n)^{1/6}/\sqrt{2}$, corresponding to the optimal bandwidth for a nonparametric kernel density estimator with Gaussian kernel (Henze and Zirkler, 1990). In addition, $d_{ij} = (\widetilde{T}_k(x_{ki}) - \widetilde{T}_k(x_{kj}))^2 + (\widetilde{T}_y(y_i) - \widetilde{T}_y(y_j))^2$, $d_i = \widetilde{T}_k^2(x_{ki}) + \widetilde{T}_y^2(y_i)$, and $\widetilde{T}_k(x_{ki})$ and $\widetilde{T}_y(y_i)$ denote the $i$th realization of $\widetilde{T}_k(X_k)$ and $\widetilde{T}_y(Y)$, respectively.

3. Select a set of important predictors with large value of $\widetilde{\omega}_k^*$, i.e., set

$$\widehat{D} = \{k : \widetilde{\omega}_k^* > cn^{-\kappa}, \text{for } \ 1 \le k \le p\},$$

where $c$ and $\kappa$ are predetermined threshold values.

Since $c$ and $\kappa$ are usually difficult to determine, we follow the other feature screening methods to set the size of $\widehat{D}$ to be $[n/\log(n)]$, where $[z]$ denotes the integer part of $z$. Since the proposed method employes the Henze-Zirkler test statistic to measure the dependence between the transformed response variable and predictors, we call it the Henze-Zirkler sure independence screening method or HZ-SIS for short.

### 2.2.2 Theoretical Properties

To study the theoretical properties of the HZ-SIS method, we first describe how the HZ-test statistic $\widetilde{\omega}_k^*$ in (2.2) is derived. Define

$$\omega_k = \int_{\mathbb{R}^2} \left| \phi_k(\mathbf{t}) - \exp(-\frac{1}{2}\mathbf{t}'\mathbf{t}) \right|^2 \varphi_\beta(\mathbf{t}) d\mathbf{t}, \quad k = 1, 2, \ldots, p,$$

where $\phi_k(\mathbf{t})$ is the characteristic function of $(\Phi^{-1}(F_k(X_k)), \Phi^{-1}(F_y(Y)))$, and $\varphi_\beta(\mathbf{t})$ is the density function of $N(0, \beta^2 I_2)$. Recall that $\exp(-\frac{1}{2}\mathbf{t}'\mathbf{t})$ is the characteristic function of $N(0, I_2)$. Therefore, $\omega_k$ can be viewed as the averaged difference between the characteristic function of the transformed variables and the characteristic function of $N(0, I_2)$. It is easy to verify that $\omega_k$ equals zero if and only if $X_k$ and $Y$ are marginally independent.

Given observations $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, where $\mathbf{x}_i = \{x_{1i}, \ldots, x_{pi}\}$ denotes the predictor variables in $i$th observation, we first use the truncated empirical distribution to estimate the CDF for each variable. In order to estimate $\omega_k$, we re-express it in the following form (Henze and Zirkler, 1990) by some algebra:

$$\begin{aligned} \omega_k = E\Big\{ & e^{-\frac{\beta^2}{2}[(\Phi^{-1}(F_k(X_k)) - \Phi^{-1}(F_k(X_k')))^2 + (\Phi^{-1}(F_y(Y)) - \Phi^{-1}(F_y(Y')))^2]} \\ & - \frac{2}{1+\beta^2} e^{-\frac{\beta^2}{2(1+\beta^2)}(\Phi^{-1}(F_k(X_k))^2 + \Phi^{-1}(F_y(Y))^2)} + \frac{1}{1+2\beta^2} \Big\}, \end{aligned}$$

where $(X_k', Y')$ is an independent copy of $(X_k, Y)$. With this representation, $\omega_k$ can be estimated using a $V$-statistic, which leads to the HZ-test statistic used in (2.2).

Next, we study the sure screening property of the HZ-SIS method. As mentioned previously, compared to the existing methods, HZ-SIS requires fewer assumptions for its sure screening property. The assumptions are given as follows.

C1 There exist positive constants $c > 0$ and $0 \leq \kappa \leq 1/4$ such that $\min_{k \in D} \omega_k \geq$

$2cn^{-\kappa}$.

C2 The dimension $p = O(\exp(n^{\tau}))$ for some constant $0 \leq \tau < \frac{1-4\kappa}{3}$.

Assumption (C1) can be viewed as a regularity condition for sure screening methods, which assumes that the minimum true signal cannot be too weak to be detectable for a given sample size, although it can gradually diminish to zero as the sample size increases to infinity. A similar assumption has been used in other methods, see e.g., Li *et al.* (2012) and Cui *et al.* (2015). Assumption (C2) allows an exponential growth of the dimension $p$ as a function of the sample size. It is also regular for ultrahigh dimensional methods. To establish the sure screening property for HZ-SIS, the key step is to establish an exponential probability bound for $|\widetilde{\omega}_k^* - \omega_k|$. The following lemma presents such an exponential probability bound with the proof given in the Appendix.

**Lemma 2.1.** *If the truncation parameter* $\delta_n = \left(4n^{\frac{m}{2}}\sqrt{2\pi m \log n}\right)^{-1}$, *where* $m = \frac{2}{3} - \frac{2\kappa}{3}$, *then there exist positive constants* $c_1 > 0$ *such that*

$$P(\max_{1\leq k\leq p} |\widetilde{\omega}_k^* - \omega_k| > cn^{-\kappa}) \leq O\left(p\exp\{-c_1 n^{\frac{1-4\kappa}{3}}\}\right),$$

*for* $1 < k \leq p$.

Here we note that we set $m = 1/2$ as the default value for the HZ-SIS method and this default value has been used in all examples of this chapter. Based on this lemma, we establish the sure screening property for HZ-SIS in the following theorem.

**Theorem 2.1.** *Under conditions (C1) and (C2), we have*

$$P(D \subseteq \hat{D}) \to 1, \quad \text{as } n \to \infty.$$

*Proof.* If $D \nsubseteq \hat{D}$, there must exist $k' \in D$ such that $\widetilde{\omega}^*_{k'} < cn^{-\kappa}$. Recall that $\omega_k > 2cn^{-\kappa}$ for all $k \in D$. Therefore, we have $|\widetilde{\omega}^*_{k'} - \omega_{k'}| > cn^{-\kappa}$. This indicates that $\{D \nsubseteq \widehat{D}\} \subseteq \{$ there exist $k' \in D$ such that $|\widetilde{\omega}^*_{k'} - \omega_{k'}| > cn^{-\kappa}\}$. Consequently,

$$\begin{aligned}
P(D \subseteq \widehat{D}) &= 1 - P(D \nsubseteq \widehat{D}) \\
&\geq 1 - P(\text{there exist } k' \in D \text{ such that } |\widetilde{\omega}^*_{k'} - \omega_{k'}| > cn^{-\kappa}) \\
&\geq 1 - P(\max_{1 \leq k \leq p} |\widetilde{\omega}^*_k - \omega_k| > cn^{-\kappa}) \geq 1 - O\{p \exp\{-c_1 n^{\frac{1-4\kappa}{3}}\}\} \\
&= 1 - O\{\exp\{-c_1 n^{\frac{1-4\kappa}{3}} + \log(p)\}\} \\
&= 1 - O\{\exp\{-c_1 n^{\frac{1-4\kappa}{3}} + c_2 n^{\tau}\}\} = 1 - o(1),
\end{aligned}$$

which concludes the proof. □

### 2.3 Simulation Studies

In this section, we used three simulated examples to assess the finite sample performance of HZ-SIS, along with comparisons with SIS (Fan and Lv, 2008), DC-SIS (Li *et al.*, 2012), Qa-SIS (He *et al.*, 2013) and MV-SIS (Cui *et al.*, 2015). In addition, NIS (Fan *et al.*, 2011) was implemented for additive model(Example 2.3.1), and the screening step in slice inversion regression for interaction detection (SIRI) was implemented for model with interactions(Example 2.3.2 & 2.3.3). For each example, we generated 100 independent datasets and summarized the performance of the methods on these 100 datasets in a few statistics. These statistics include the minimum size of $\widehat{D}$ needed to cover all active variables, which is denoted by MSD for short; and for the given size $\nu_n = [n/\log(n)]$ of $\widehat{D}$, the proportion of $\widehat{D}$ covering a single active predictor $X_k$ (denoted by $P_k$), and the proportion of $\widehat{D}$ covering all active variables (denoted by $P_a$). The reason for choosing the above statistics is that in practice, we usually specify the size $\nu_n$ of $\widehat{D}$ instead of the thresholding value $cn^{-\kappa}$

for feature screening.

### 2.3.1 An Additive Model Example

This example is adopted from Cui *et al.* (2015). Let

$$f_1(x) = -\sin(2x), \quad f_2(x) = x^2 - \frac{12}{25}, \quad f_3(x) = x, \quad f_4(x) = e^{-x} - \frac{2}{5}\sinh(\frac{5}{2}),$$

and consider the additive model

$$Y = 3f_1(X_1) + f_2(X_2) - 1.5f_3(X_3) + f_4(X_4) + \varepsilon,$$

where the error term $\varepsilon$ follows a $t(1)$ distribution. For the predictors, we consider two different distributions:

1. $X_k$'s, $k = 1, \ldots, p$, are generated independently from the distribution $t(4)$;
2. $X_k$'s, $k = 1, \ldots, p$, are generated independently from the Uniform[$-2.5$,2.5].

We set $(n, p) = (200, 2000)$ and repeated each case for 100 times. In Qa-SIS, we set $\tau = 0.5$ and the number of basis $d_n = [n^{\frac{1}{5}}] = 3$. In NIS, we took the number of basis $d_n = [n^{\frac{1}{5}}] + 2 = 5$. In MV-SIS, we discretized each predictor into a four-categorical variable using the first, second and third quartiles as knots. For MV-SIS, the same discretization method has been used in all examples of this paper. The results are summarized in Table 2.1.

From Table 2.1, we can see that when the predictors are generated from $t(4)$, a heavy-tailed distribution, HZ-SIS performs best, followed by MV-SIS, DC-SIS and Qa-SIS. This result, combined with the fact that HZ-SIS requires fewer assumptions for the sure screening property, indicates that HZ-SIS is a more robust feature screening method than the existing ones. When the predictors are generated from

Table 2.1: Simulation results for the additive model example. For MSD, we report the median with its associated interquartile range (IQR) in the parentheses.

| | Method | MSD | $P_1$ | $P_2$ | $P_3$ | $P_4$ | $P_a$ |
|---|---|---|---|---|---|---|---|
| Case 1 | SIS | 976.50(1023.00) | 0.07 | 0.08 | 0.96 | 0.98 | 0.01 |
| | NIS | 1342.50(704.75) | 0.01 | 0.20 | 0.05 | 0.76 | 0.00 |
| | DC-SIS | 279.50(656.75) | 0.27 | 0.46 | 0.57 | 0.95 | 0.16 |
| | MV-SIS | 24.00(118.00) | 0.83 | 0.73 | 0.97 | 0.95 | 0.58 |
| | Qa-SIS | 347.50(653.50) | 0.02 | 0.81 | 0.22 | 0.98 | 0.00 |
| | HZ-SIS | 11.50(22.00) | 0.98 | 0.90 | 0.97 | 0.94 | 0.80 |
| Case 2 | SIS | 1216.50(964.75) | 0.10 | 0.02 | 1.00 | 1.00 | 0.00 |
| | NIS | 924.00(1257.25) | 0.16 | 0.17 | 0.30 | 0.33 | 0.06 |
| | DC-SIS | 197.00(339.00) | 0.20 | 0.31 | 0.98 | 0.98 | 0.06 |
| | MV-SIS | 11.00(28.00) | 0.94 | 0.83 | 1.00 | 1.00 | 0.78 |
| | Qa-SIS | 8.00(15.50) | 0.91 | 0.91 | 1.00 | 1.00 | 0.82 |
| | HZ-SIS | 24.50(52.25) | 0.71 | 0.88 | 0.99 | 1.00 | 0.64 |

the uniform distribution, for which the support is bounded, HZ-SIS still performs reasonably well. In this case, it is comparable with MV-SIS and Qa-SIS, but much better than DC-SIS, NIS and SIS.

For the case where the predictors are generated from $t(4)$ distribution, we also plotted histograms of the calculated screening indices of each method. Specifically, for each method, we first combined the corresponding screening indices from 100 simulations. Then we drew a histogram using all 400 indices from active variables and a histogram using 600 indices from inactive variables, which are randomly selected from a total of 199,600($100 \times 1996$) ones. Finally, we put two histograms in the same figure and differentiated them by color. The histograms are shown in Figure 2.1. It is clear that HZ-SIS has the smallest overlapping area for its two histograms, which again confirms its superiority in separating active features and inactive ones.

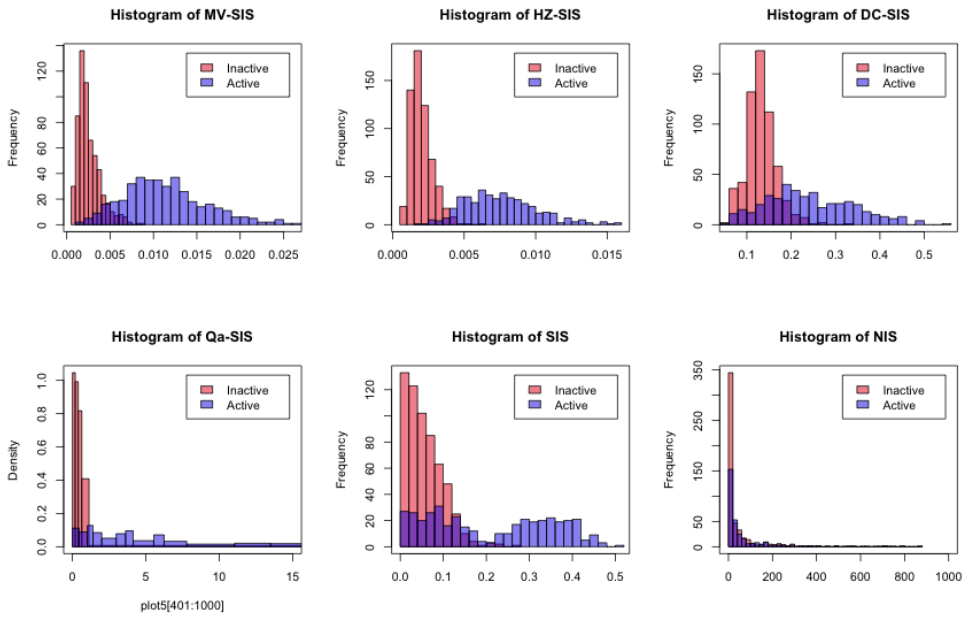


Figure 2.1: Histograms of the screening indices of different methods for the additive model example with the predictors generated from the distribution $t(4)$.

### 2.3.2 A Model with Interaction Variables

This example illustrates the performance of HZ-SIS for the models with interaction variables. Let

$$Y = 0.5 + \frac{10X_1}{1 + X_{50}^2} + \varepsilon,$$

The vector of covariates $\mathbf{X} = (X_1, \cdots, X_p)^T$ is generated from the multivariate normal distribution having mean $\mathbf{0}$ and the covariance matrix $\mathbf{\Sigma} = (\sigma_{ij})_{p\times p}$ with $\sigma_{ij} = 0.5^{|i-j|}$. For the error term $\varepsilon$, we considered two cases: (i) $\epsilon$ follows $N(0, 1^2)$ distribution. (ii) $\epsilon$ follows $t(1)$ distribution. We set $(n, p) = (200, 1000)$ and repeated each experiment for 100 times.

Jiang and Liu (2014) recently proposed a procedure, called sliced inverse regression for interaction detection (SIRI), to conduct high dimensional variable selection

for the model with interaction terms. Instead of building a predictive model of the response given combinations of predictors, this procedure is based on modeling the conditional distribution of predictors given responses. Since this procedure includes a screening step, so we also implemented this step here and denoted it as SIRI-SIS.

In SIRI-SIS, we used a fixed slicing scheme with 10 slices of equal size (H=10). In Qa-SIS procedure, we set $\tau = 0.4$ and the number of basis $d_n = [n^{\frac{1}{5}}] = 3$. The results are summarized in Table 2.2.

Table 2.2: Results for the model with interaction variables. For MSD, we report the median with its associated interquartile range (IQR) in the parentheses.

| | Method | MSD | $P_1$ | $P_{50}$ | $P_a$ |
|---|---|---|---|---|---|
| Case 1 | SIS | 686.00(321.25) | 1.00 | 0.00 | 0.00 |
| | SIRI-SIS | 3.00(1.00) | 1.00 | 0.99 | 0.99 |
| | DC-SIS | 50.50(52.25) | 1.00 | 0.39 | 0.39 |
| | MV-SIS | 34.00(73.75) | 1.00 | 0.52 | 0.52 |
| | Qa-SIS | 426.00(478.00) | 1.00 | 0.02 | 0.02 |
| | HZ-SIS | 3.00(1.00) | 1.00 | 1.00 | 1.00 |
| Case 2 | SIS | 575.50(387.00) | 0.78 | 0.03 | 0.01 |
| | SIRI-SIS | 11.00(52.00) | 1.00 | 0.69 | 0.69 |
| | DC-SIS | 167.00(251.25) | 1.00 | 0.12 | 0.12 |
| | MV-SIS | 97.50(158.75) | 1.00 | 0.23 | 0.23 |
| | Qa-SIS | 414.50(403.75) | 1.00 | 0.02 | 0.02 |
| | HZ-SIS | 9.50(22.00) | 1.00 | 0.86 | 0.86 |

Table 2.2 indicates that in the case where error term is normal, all methods can detect $X_1$ with ease, but when it comes to detecting $X_{50}$, HZ-SIS and SIRI-SIS substantially outperforms other methods. For the case where error term follows $t(1)$ distribution, we have similar conclusions as in the normal case. In addition, our method performs slightly better than SIRI-SIS in this case.

To understand the performance of these methods, we show in Figure 2.2 the scat-

ter plots of the transformed predictors $\widetilde{T}_1(X_1)$, $\widetilde{T}_{50}(X_{50})$ and $\widetilde{T}_{100}(X_{100})$ versus the transformed response variable $\widetilde{T}_y(Y)$ in case 1. The scatter plots of $X_1$, $X_{50}$ and $X_{100}$ versus $Y$ are similar. Given the reference scatter plot of $(\widetilde{T}_{100}(X_{100}), \widetilde{T}_y(Y))$ for which the theoretical joint distribution is $N(0, I_2)$, we can see that the joint distributions of $(\widetilde{T}_1(X_1), \widetilde{T}_y(Y))$ and $(\widetilde{T}_{50}(X_{50}), \widetilde{T}_y(Y))$ substantially deviate from $N(0, I_2)$, and thereby HZ-test is powerful in detecting the dependence of $Y$ on $X_1$ and $X_{50}$. However, not all other methods work well for this example. As indicated by the values of $P_2$ reported in Table 2.2, SIS and Qa-SIS essentially fail to detect the dependence of $Y$ on $X_{50}$, and DC-SIS and MV-SIS have only limited success probabilities of detecting this dependence.

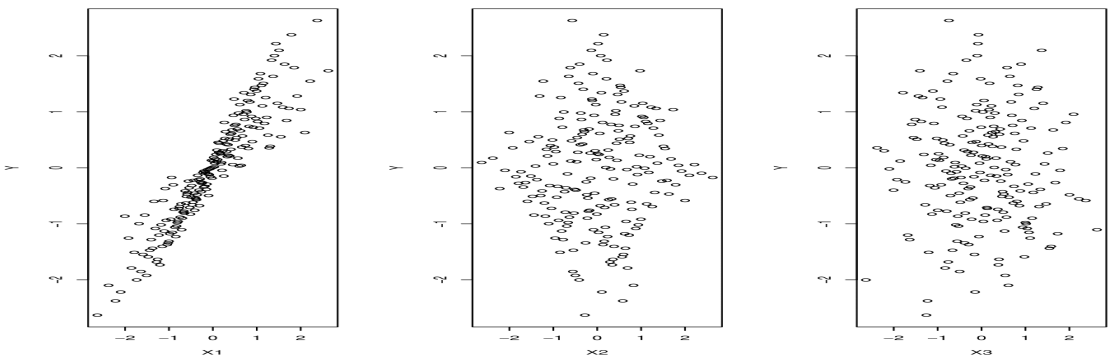


Figure 2.2: Scatter plots of the transformed response variable $\widetilde{T}_y(Y)$ versus the transformed predictors $\widetilde{T}_1(X_1)$, $\widetilde{T}_{50}(X_{50})$ and $\widetilde{T}_{100}(X_{100})$

### 2.3.3 A Complex Model with More Interaction Variables

This example illustrates the performance of HZ-SIS for more complex models. Let

$$Y = 1 + A[10X_1 + \exp(X_2^2 + 3X_3)] + 10\frac{X_5}{2 + X_6} + 3(X_7 + X_8)^2 + \varepsilon,$$

Table 2.3: Results for the complex model with more interaction variables. For MSD, we report the median with its associated interquartile range (IQR) in the parentheses.

| | Method | MSD | $P_1$ | $P_2$ | $P_5$ | $P_6$ | $P_7$ | $P_8$ | $P_a$ |
|---|---|---|---|---|---|---|---|---|---|
| Case1 | SIS | 841.50(196.25) | 0.07 | 0.30 | 0.78 | 0.06 | 0.08 | 0.06 | 0.00 |
| | SIRI-SIS | 470.00(377.25) | 0.73 | 0.99 | 0.34 | 0.05 | 0.48 | 0.53 | 0.00 |
| | DC-SIS | 911.00(117.00) | 0.04 | 0.94 | 0.04 | 0.07 | 0.04 | 0.07 | 0.00 |
| | MV-SIS | 511.50(322.00) | 0.28 | 1.00 | 0.88 | 0.09 | 0.38 | 0.33 | 0.00 |
| | Qa-SIS | 420.50(305.25) | 0.02 | 0.98 | 0.09 | 0.05 | 0.05 | 0.08 | 0.00 |
| | HZ-SIS | 245.50(341.75) | 0.90 | 1.00 | 0.93 | 0.30 | 0.76 | 0.53 | 0.12 |
| Case2 | SIS | 882.50(183.50) | 0.10 | 0.16 | 0.84 | 0.07 | 0.09 | 0.11 | 0.00 |
| | SIRI-SIS | 432.00(378.25) | 0.56 | 1.00 | 0.17 | 0.06 | 0.45 | 0.40 | 0.00 |
| | DC-SIS | 899.50(134.25) | 0.10 | 0.11 | 0.10 | 0.09 | 0.03 | 0.08 | 0.00 |
| | MV-SIS | 484.00(325.00) | 0.63 | 1.00 | 0.85 | 0.11 | 0.39 | 0.36 | 0.00 |
| | Qa-SIS | 416.00(453.25) | 0.04 | 0.99 | 0.06 | 0.08 | 0.03 | 0.08 | 0.00 |
| | HZ-SIS | 200.50(309.25) | 0.79 | 1.00 | 0.89 | 0.26 | 0.71 | 0.40 | 0.10 |

where $A$ is generated from the set $\{-1, 1\}$ with equal probability, $X_k$'s are independently generated from $t(4)$ distribution. For the error term $\varepsilon$, we considered two cases: (i) $\epsilon$ follows $N(0, 1^2)$ distribution. (ii) $\epsilon$ follows $t(1)$ distribution. This model is complex, containing more interaction variables than previous examples. We set $(n, p) = (400, 1000)$ and repeated each experiment 100 times.

In SIRI-SIS, we used a fixed slicing scheme with 10 slices of equal size (H=10). In Qa-SIS procedure, we set $\tau = 0.4$ and the number of basis $d_n = [n^{\frac{1}{5}}] = 3$. The results are summarized in Table 2.3.

From Table 2.3, we can see that in both case, HZ-SIS has an overall superior performance against the other methods.

## 2.4 Screening of Anticancer Drug Response Genes

Recent advances in high-throughput biotechnologies, such as microarray, sequencing technologies and mass spectrometry, have provided an unprecedented opportunity for biomarker discovery. Molecular biomarkers can not only facilitate disease diagno-

sis, but also reveal underlying, biologically distinct, patient subgroups with different sensitivities to a specific therapy. The latter is known as disease heterogeneity, which is often observed in complex diseases such as cancer. For example, molecularly targeted cancer drugs are only effective for patients with tumors expressing targets (Grünwald and Hidalgo, 2003; Buzdar, 2009). The disease heterogeneity has directly motivated the development of precision medicine, which aims to improve patient care by tailoring optimal therapies to an individual patient according to his/her molecular profile and other clinical characteristics.

Toward the ultimate goal of precision medicine, i.e., selecting right drugs for individual patients, a recent large-scale pharmacogenomics study, namely, cancer cell line encyclopedia (CCLE), has screened multiple anticancer drugs over hundreds of cell lines in order to elucidate the response mechanism of anticancer drugs. The dataset consists of the dose-response data for 24 chemical compounds across over 479 cell lines. For each cell line, it consists of the expression data of 18,988 genes. The dataset is publicly available at *www.broadinstitute.org/ccle*. Our goal is to screen the genes that respond to each chemical compounds, which will facilitate the followed analysis for identification of anticancer drug response genes. In our analysis, we used the area under the dose-response curve, which is termed as activity area in Barretina *et al.* (2012), to measure the sensitivity of drug to a given cell line. Compared to other measurements, such as $IC_{50}$ and $EC_{50}$, the activity area could capture the efficacy and potency of a drug simultaneously.

The drug topotecan (trade name Hycamtin) is a chemotherapeutic agent that is a topoisomerase inhibitor. It is a synthetic, water-soluble analog of the natural chemical compound camptothecin and has been used to treat ovarian cancer, lung cancer and other cancer types. After GlaxoSmithKline received final FDA approval for Hycamtin Capsules in 2007, topotecan became the first topoisomerase I inhibitor

Table 2.4: Top 10 genes selected for the drug topotecan by different methods.

| Rank | SIS | DC-SIS | MV-SIS | Qa-SIS | HZ-SIS |
|---|---|---|---|---|---|
| 1 | CADM2 | ITGB5 | HMGB2 | FLJ35816 | RFXAP |
| 2 | MMP27 | PPIC | KIF15 | GATS | HMGB2 |
| 3 | WNT5B | HMGB2 | RFXAP | HS3ST3A1 | ITGB5 |
| 4 | CDX4 | ARHGAP19 | ARHGAP19 | ASIC4 | BCLAF1 |
| 5 | ELF4 | RFXAP | CD63 | TRAV26-2 | CPSF6 |
| 6 | ECI2 | CPSF6 | TAF5 | LOC10 | HAUS1 |
| 7 | GLIPR1 | CD63 | CPSF6 | LOC11 | ILF3 |
| 8 | ABCC9 | TAF5 | ELAVL1 | VPS72 | ELAVL1 |
| 9 | ADCY5 | CNTRL | ILF3 | PIK3IP1 | TAF5 |
| 10 | PPIC | SLFN11 | S100A10 | CRSF6 | SLFN11 |

for oral use. Table 2.4 lists the top 10 important genes selected for topotecan by HZ-SIS. For comparison, the table also includes the top 10 genes selected by SIS, DC-SIS, MV-SIS and Qa-SIS. In Qa-SIS procedure, we set $\tau = 0.5$ and the number of basis $d_n = [n^{\frac{1}{5}}] = 3$. For topotecan, the gene SLFN11 has been recognized as a very important predictor for the sensitivity of topotecan (Barretina *et al.*, 2012; Zoppoli *et al.*, 2012). HZ-SIS ranks it No. 10. In addition to SLFN11, Wang *et al.* (2014) found the strong relevance of HMGB2 and BCLAF1 to topotecan. HZ-SIS ranks these two genes No. 2 and No. 4, respectively. DC-SIS has a similar performance to HZ-SIS for the drug topotecan, while the other methods do not.

The drug 17-AAG is a derivative of the antibiotic geldanamycin that is being studied in the treatment of cancer, specific young patients with certain types of leukemia or solid tumors, especially kidney tumors. 17-AAG works by inhibiting the gene HSP90, which is expressed in those tumors, and belongs to the family of drugs called antitumor antibiotics. Table 2.5 reports the top 10 genes ranked by different methods for 17-AAG. According to Hadley and Hendricks (2014) and Barretina *et al.* (2012), the gene NQO1 is the top predictive biomarker for 17-AAG. HZ-SIS ranks it

Table 2.5: Top 10 genes selected for the drug 17-AAG by different methods.

| Rank | SIS | DC-SIS | MV-SIS | Qa-SIS | HZ-SIS |
|---|---|---|---|---|---|
| 1 | UXT | NQO1 | NQO1 | MMP24 | NQO1 |
| 2 | IGFN1 | MMP24 | INO80 | ATP6V0E1 | OGDHL |
| 3 | MSH2 | ZNF610 | MMP24 | ZFP30 | TMEM198 |
| 4 | ROCK1 | ZFP30 | ZNF610 | GPR35 | ZBTB7A |
| 5 | DDA1 | NFKB1 | ZFP30 | SLC1A5 | GYG2 |
| 6 | SCEL | CDH6 | PRPUSD4 | GPX2 | CDH6 |
| 7 | ST5 | OGDHL | LOC10 | CNTRL | ZNF610 |
| 8 | THUMPD3 | LOC10 | PCSK1N | VPS72 | RPUSD4 |
| 9 | ITGA9 | PRUSD4 | NFKB1 | LOC10 | CSK |
| 10 | C20orf141 | IN080 | ZBTB7A | ZNF610 | CTCF |

first among all genes. DC-SIS and MV-SIS also rank it first.

# 3. DOUBLE-PARALLEL MONTE CARLO FOR BAYESIAN ANALYSIS OF BIG DATA

## 3.1 Introduction

The MCMC method has proven to be a very powerful and typically unique computational tool for analyzing data of complex structures. However, it is difficult to be applied to big data problems for which complex models are often needed. The difficulty comes from two aspects. The first one is on data storage; the dataset can be too large for a single computer to store and process. The second one is on computational time; the MCMC method can be very time consuming for simulating from the posterior of a large data set, which typically requires a large number of iterations and a complete scan of the full dataset for each iteration. However, thanks to strategy of embarrassingly parallel computing, the two issues can now be solved simultaneously.

The strategy of embarrassingly parallel computing is to divide a large dataset into a number of smaller subsets such that each subset can be stored in a single machine, and then conduct the Bayesian analysis for each subset separately. Finally, the posterior samples generated for each subset are aggregated in some way such that a correct inference can be made for the full data posterior. During the past few years, this strategy has been pursued by a few groups enthusiastically. Several algorithms have been developed to address the issue of subset posterior aggregation.

To be a little more detailed, suppose that a large dataset has been partitioned into $k$ subsets, and $N$ posterior samples have been generated for each subset. Let $\{\boldsymbol{\theta}_1^{(i)}, \ldots \boldsymbol{\theta}_N^{(i)}\}$ denote the posterior samples generated from subset $i$. Based on the Bernstein-von Mises theorem, which states that the posterior tends to a normal

distribution centered around the true parameter value $\boldsymbol{\theta}^*$ as the number of observations grows, Scott *et al.* (2016) proposed to use the weighted average $\sum_{i=1}^{k} w_i \boldsymbol{\theta}_j^{(i)}$ to approximate a full data posterior sample, where the weight $w_i$ is the inverse of the covariance matrix of $\{\boldsymbol{\theta}_1^{(i)}, \ldots \boldsymbol{\theta}_N^{(i)}\}$. This algorithm is exact when the subset posterior is exactly Gaussian. Based on the same theory, Neiswanger *et al.* (2013) proposed to fit the posterior samples generated for each subset by a Gaussian density, denoted the fitted density by $\hat{p}_i$ for $i = 1, \ldots, k$, and then draw samples from the product density $\hat{p}_1 \ldots \hat{p}_k$. As an extension of this approach, Neiswanger *et al.* (2013) also proposed to estimate the subset posterior density using a Gaussian kernel density estimation method or a semiparametric density estimation method. Wang and Dunson (2013) proposed a Weierstrass refinement sampler, where the samples from an initial approximation to the full data posterior (e.g., obtained via variational approximation or other methods) are refined using the information obtained from the subset posterior samples within a Weierstrass approximation. Another method that makes use of kernel approximation is by Minsker *et al.* (2014), where the subset posteriors are combined by estimating a probability distribution that minimizes a loss function defined in the reproducing kernel Hilbert space embedding the subset posteriors. These methods generally work well, but their accuracy can vary significantly depending on how close the subset posteriors are to Gaussian or the choice of kernel and its bandwidth. In particular, their accuracy can be low when the dimension of $\boldsymbol{\theta}$ is high. Quite recently, the so-called WASP method was proposed by Srivastava *et al.* (2015), where each subset posterior is approximated by an empirical measure and they are combined by estimating their barycenter in the Wasserstein space of probability measures. This method does not depend on the kernel density estimation any more, but computing the Wasserstein barycenter needs to solve a huge linear programming problem which often requires a lot of computer memory.

In this chapter, we introduce a new method for aggregating subset posterior samples. The new method is surprisingly simple, which is to first simulate from some modified subset posteriors, for which the log-likelihood functions are appropriately scaled according to their sample size, and then recenter the subset posterior samples to their global mean. Under mild conditions, we show that the aggregated samples have the same convergence rate toward the true parameter $\boldsymbol{\theta}^*$ as those drawn from the full data posterior. The numerical results indicate that the new method can be rather accurate compared to the existing ones. In order to further speed up computation, we suggest to use the Pop-SAMC algorithm (Song *et al.*, 2014), rather than traditional single chain MCMC algorithms, to draw samples from each subset posterior. Since the proposed method consists of two levels of parallel, data parallel and simulation parallel, it is coined as "double parallel" Monte Carlo.

The remainder of this chapter is organized as follows. Section 3.2 presents the proposed sample aggregation method and describes its theoretical properties. Section 3.3 first gives a brief review of the pop-SAMC algorithm, and then discusses the double parallel strategy. Sections 3.4 and 3.5 present some numerical results along with some comparisons with the existing methods.

## 3.2 Subset Posterior Aggregation

Suppose that a random sample $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ has been collected from the distribution $f(x|\boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}^* \in \Theta \subset \mathbb{R}^p$ and $\Theta$ is the parameter space. Let $g(\boldsymbol{\theta})$ denote the prior distribution of $\boldsymbol{\theta}$. Then the posterior distribution of $\boldsymbol{\theta}$ is given by

$$\pi(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{\prod_{i=1}^n f(X_i|\boldsymbol{\theta})g(\boldsymbol{\theta})}{\int_\Theta \prod_{i=1}^n f(X_i|\boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{3.1}$$

In most cases, $\pi(\boldsymbol{\theta}|\boldsymbol{X})$ is analytically intractable and we have to approximate it using the Markov chain Monte Carlo(MCMC) method. However, as mentioned previously,

when $n$ is very large, the MCMC method is computationally prohibitive because it requires a large number of scans of the dataset.

To address this issue, we divide the data into k subsets, each containing about the same number of samples. Let $\boldsymbol{X}_{[j]} = (X_{j1}, \ldots, X_{jm_j})$ denote the $j$th subset, where $m_j$ denote the sample size of $\boldsymbol{X}_{[j]}$. Let $\pi(\boldsymbol{\theta}|\boldsymbol{X}_{[j]})$ denote the posterior distribution corresponding to the subset $\boldsymbol{X}_{[j]}$, for which the variance is approximately $n/m_j$ times the variance of full data posterior $\pi(\boldsymbol{\theta}|\boldsymbol{X})$. To adjust the variance, for each subset, we instead work on a modified subset posterior

$$\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{X}_{[j]}) = \frac{\prod_{i=1}^{m} f^{n/m_j}(X_{ji}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\Theta} \prod_{i=1}^{m} f^{n/m_j}(X_{ji}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}, \tag{3.2}$$

where each sample is duplicated $n/m_j$ times. Such a modification, first introduced in Minsker *et al.* (2014), ensures that $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{X}_{[j]})$ has about the same variance as the full data posterior. In what follows, we refers to $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{X}_{[j]})$ as a subposterior and, without loss of generality, assume that $m_1 = m_2 = \cdots = m_k = m$ holds.

Let $\boldsymbol{\mu}^{(1)}, \cdots, \boldsymbol{\mu}^{(k)}$ denote the mean of the subposteriors, and let $\hat{\boldsymbol{\mu}} = \frac{1}{k}\sum_{j=1}^{k} \boldsymbol{\mu}^{(j)}$ denote their averages. We propose to recenter each of the subposteriors to $\hat{\boldsymbol{\mu}}$ and then use the following mixture of re-centered subposteriors to approximate the full data posterior $\pi(\boldsymbol{\theta}|\boldsymbol{X})$:

$$\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{1}{k}\sum_{j=1}^{k} \tilde{\pi}(\boldsymbol{\theta} + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^{(j)})|\boldsymbol{X}_{[j]}). \tag{3.3}$$

To quantify the accuracy of the approximation, we make the following assumptions:

(A1) The log-likelihood function $\sum_{i=1}^{m} \log f(x_{ji}|\boldsymbol{\theta})$ is Laplace-regular for each $j = 1, \ldots, k$.

(A2) $\boldsymbol{\theta}^*$ is an interior point of $\Theta$, $g(\boldsymbol{\theta}^*) > 0$, and $g(\boldsymbol{\theta})$ is four times continuous differentiable on $\Theta$.

(A3) The number of subsets $k$ can increase slowly with $n$, but can not exceed $O(n^{1/2})$.

Since the quantification involves posterior expansions based on Laplace's method, the Laplace regularity condition is assumed. Refer to Kass *et al.* (1990) for the detail. This condition is standard and generally holds for the exponentially family. Under the above conditions, we have the following theorem, whose proof is given in the appendix.

**Theorem 3.1.** *If the conditions (A1)-(A3) are satisfied, then we have*

$$E[E_{\tilde{\pi}}(\boldsymbol{\theta}) - E_{\pi}(\boldsymbol{\theta})]^2 = O(m^{-2}), \tag{3.4}$$

$$E|\,Var_{\tilde{\pi}}(\boldsymbol{\theta}) - Var_{\pi}(\boldsymbol{\theta})| = o(n^{-1}), \tag{3.5}$$

$$E(d^2(\pi, \delta_{\boldsymbol{\theta}^*})) = 2\frac{tr(I^{-1})}{n} + o(n^{-1}), \tag{3.6}$$

$$E(d^2(\tilde{\pi}, \delta_{\boldsymbol{\theta}^*})) = 2\frac{tr(I^{-1})}{n} + o(n^{-1}), \tag{3.7}$$

*where $E_{\pi}$ and $E_{\tilde{\pi}}$ denote the expectations with respect to $\pi$ and $\tilde{\pi}$, respectively; $Var_{\pi}$ and $Var_{\tilde{\pi}}$ denote the variances with respect to $\pi$ and $\tilde{\pi}$, respectively; $I = -E_{X|\boldsymbol{\theta}^*}\frac{\partial^2 \log f(X|\boldsymbol{\theta}^{(*)})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}$ is the Fisher information matrix, and $d^2(\tilde{\pi}, \delta_{\boldsymbol{\theta}^*}) = \int_{\Theta} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{X})d\boldsymbol{\theta}$ is the Wasserstein distance of order 2 between $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{X})$ and the Dirac measure at $\boldsymbol{\theta}^*$.*

Equations (3.4) and (3.5) measure the accuracy of the approximation $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{X})$ to $\pi(\boldsymbol{\theta}|\boldsymbol{X})$ in terms of mean and variance, respectively. In particular, equation (3.4) implies that $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{X})$ and $\pi(\boldsymbol{\theta}|\boldsymbol{X})$ will lead to the same Bayesian estimate (with respect to the square loss function), and equation (3.5) implies that the Bayesian

estimates led from $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{X})$ and $\pi(\boldsymbol{\theta}|\boldsymbol{X})$ will have about the same variance when $n$ is large. Equations (3.6) and (3.7) imply that $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{X})$ and $\pi(\boldsymbol{\theta}|\boldsymbol{X})$ share the same convergence rate toward the true value $\boldsymbol{\theta}^*$. In other words, the subposterior aggregation does not lose much information about the data.

Rather than $\boldsymbol{\theta}$ itself, sometimes we are interested in $h(\boldsymbol{\theta})$, a $\mathbb{R}^p \mapsto \mathbb{R}^q$ function of $\boldsymbol{\theta}$. A similar result, which measures the accuracy of the approximation $\tilde{\pi}(h(\boldsymbol{\theta})|\boldsymbol{X})$, can be obtained under the following condition:

(A4) $h(\boldsymbol{\theta})$ is square integrable and thrice times continuous differentiable in a neighborhood of $\boldsymbol{\theta}^*$.

**Corollary 3.1.** *If A1-A4 are satisfied, then we have*

$$\begin{aligned}
E[E_{\tilde{\pi}}h(\boldsymbol{\theta}) - E_{\pi}h(\boldsymbol{\theta})]^2 &= O(m^{-2}), \\
E|Var_{\tilde{\pi}}h(\boldsymbol{\theta}) - Var_{\pi}h(\boldsymbol{\theta})| &= o(n^{-1}), \\
E(d^2(\pi(h(\boldsymbol{\theta})|\boldsymbol{X}), \delta_{h(\boldsymbol{\theta}^*)})) &= 2\frac{tr(H^*_{(1)}I^{-1}H^{*'}_{(1)})}{n} + o(n^{-1}), \\
E(d^2(\tilde{\pi}(h(\boldsymbol{\theta})|\boldsymbol{X}), \delta_{h(\boldsymbol{\theta}^*)})) &= 2\frac{tr(H^*_{(1)}I^{-1}H^{*'}_{(1)})}{n} + o(n^{-1}),
\end{aligned}$$

*where* $H^*_{(1)} = \frac{\partial h(\theta)}{\partial \boldsymbol{\theta}^T}|_{\theta=\theta^*}$*, and* $I$ *is the Fisher information matrix as defined in Theorem 3.1.*

The proof is similar to that of Theorem 3.1, which is based on the expansion for the posterior mean of $h(\boldsymbol{\theta})$ and thus omitted here.

### 3.3 Double Parallel Monte Carlo

In this section, we first give a brief review of the Pop-SAMC algorithm and discuss its implementation on the OpenMP platform. Then we describe the double parallel Monte Carlo scheme.

### 3.3.1 Pop-SAMC Algorithm and Its OpenMP Implementation

As aforementioned, although MCMC is powerful for analyzing the data of complex structures, its computer-intensive nature precludes its use for big data analysis. To accelerate computation, one feasible way is to conduct parallel MCMC simulations. People have debated for a long time to make a single long run or many short runs. For conventional MCMC algorithms, such as the Metropolis-Hastings algorithm(Metropolis *et al.*, 1953; Hasting, 1970) and the Gibbs sampler (Geman and Geman, 1984), parallel runs may not provide any theoretical advantages over a single long run. In general, if you cannot get a good answer with a long run, then you cannot get a good answer with many short runs either. However, this situation differs for the population stochastic approximation Monte Carlo (pop-SAMC) algorithm (Song *et al.*, 2014), where it is shown that running pop-SAMC with $\kappa$ chains (in parallel) for $T$ iterations is asymptotically more efficient than running a single SAMC chain for $\kappa T$ iterations when the gain factor sequence decreases slower than $O(1/t)$, where $t$ indexes iterations. This is due to that the chains in pop-SAMC interact with each other intrinsically.

The pop-SAMC algorithm consists of two steps, population sampling and $\boldsymbol{\xi}$-updating, where $\boldsymbol{\xi}$ denotes an adaptive parameter evolving with iterations. In the population sampling step, each chain is updated independently for one or a few iterations. In the $\boldsymbol{\xi}$-updating step, $\boldsymbol{\xi}_t$ (i.e., the value of $\boldsymbol{\xi}$ at iteration $t$) is updated based on the collected information from individual chains, which enforces interactions between different chains and, consequently, improves the efficiency of the algorithm. The detailed algorithm is described below.

Suppose that we are interested in simulating samples from a density function $p(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, and $\Theta$ has been partitioned into $M$ subregions: $E_1 = \{\boldsymbol{\theta} : U(\boldsymbol{\theta}) <$

$u_1\}$, $E_2 = \{\boldsymbol{\theta} : u_1 \le U(\boldsymbol{\theta}) < u_2\}$, ..., $E_{M-1} = \{\boldsymbol{\theta} : u_{M-2} \le U(\boldsymbol{\theta}) < u_{M-1}\}$, and $E_M = \{\boldsymbol{\theta} : U(\boldsymbol{\theta}) \ge u_{M-1}\}$, where $U(\boldsymbol{\theta})$ is a pre-specified function of $\boldsymbol{\theta}$, e.g., $U(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta})$, and $u_1 < u_2 < \cdots < u_{M-1}$ are pre-specified numbers. To explain the concept of SAMC, we assume for the time being that all the subregions are non-empty; that is, $z_i = \int_{E_i} p(\boldsymbol{\theta}) d\boldsymbol{\theta} > 0$ for all $i = 1, \ldots, M$. However, as explained in Liang *et al.* (2007), the algorithm does allow the existence of empty subregions. Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$ denote the desired sampling distribution of the $M$ subregions, where $\sum_{i=1}^M \pi_i = 1$ and $\pi_i > 0$ for all $i = 1, \ldots, M$. Given the partition and the desired sampling distribution, Pop-SAMC seeks to draw samples from the distribution

$$p_{\boldsymbol{z}}(\boldsymbol{\theta}) \propto \sum_{i=1}^{M} \frac{\pi_i p(\boldsymbol{\theta})}{z_i} I(\boldsymbol{\theta} \in E_i).$$

If $z_i$'s are known and the space is partitioned appropriately, e.g., the energy bandwidth of each subregion is small enough, then the sampling will lead to a random walk in the space of subregions and thus the local-trap problem can be overcome essentially. However, since $z_1, \ldots, z_M$ are generally unknown, Pop-SAMC employs the stochastic approximation algorithm (Robbins and Monro, 1951) to learn their values (up to a constant factor) in an adaptive way.

Let $\kappa$ denote the population size, i.e., the number of parallel Markov chains contained in Pop-SAMC, and let $\boldsymbol{\theta}_t = (\boldsymbol{\theta}_{t1}, \ldots, \boldsymbol{\theta}_{t\kappa})$ denote the current state of the $\kappa$ chains. Let $\boldsymbol{\xi}_t = (\xi_{t1}, \ldots, \xi_{tM})$ denote the working estimate of $(z_1/\pi_1, \ldots, z_M/\pi_M)$ obtained at iteration $t$. One iteration of the algorithm consists of the following steps:

1. (Population sampling) For $i = 1, \ldots, \kappa$, generate a new sample $\boldsymbol{\theta}_{t,i}$ starting from $\boldsymbol{\theta}_{t-1,i}$ by a single MH update with the target distribution given by

$$p_{\boldsymbol{\xi}_{t-1}}(\boldsymbol{\theta}) \propto \sum_{j=1}^{M} \frac{p(\boldsymbol{\theta})}{e^{\xi_{t-1,j}}} I(\boldsymbol{\theta} \in E_j). \tag{3.8}$$

2. ($\boldsymbol{\xi}$-update) Set $\boldsymbol{\xi}_t = \boldsymbol{\xi}_{t-1} + \gamma_t(\boldsymbol{H}_t - (1/M)\mathbf{1})$, where $\boldsymbol{H}_t = (\sum_{i=1}^{\kappa} I(\boldsymbol{\theta}_{t,i} \in E_1)/\kappa, \cdots, \sum_{i=1}^{N} I(\boldsymbol{\theta}_{t,i} \in E_M)/\kappa)^T$, and $\gamma_t$ is a gain factor.

To ensure the convergence of the algorithm, the gain factor $\{\gamma_t\}$ is required to satisfy the conditions:

$$\sum_{t=1}^{\infty} \gamma_t = \infty, \quad \frac{\gamma_{t+1} - \gamma_t}{\gamma_t} = O(\gamma_{t+1}^{\tau}), \quad \sum_{t=1}^{\infty} \frac{\gamma_t^{(1+\tau')/2}}{\sqrt{t}} < \infty,$$

for some $\tau \in [1, 2)$ and $\tau' \in (0, 1)$. For example, one can set $\gamma_t = O(1/t^{\zeta})$ for $\zeta \in (1/2, 1]$. To accommodate the case that $\boldsymbol{\xi}_t$ takes values in an unbounded space, a varying truncation version of the algorithm can be considered as in Andrieu *et al.* (2005).

Like the SAMC algorithm(Liang *et al.*, 2007), Pop-SAMC possesses the self-adjusting mechanism, which operates based on past samples and enables the simulation to be immune to local traps. This can be considered as a significant advantage over conventional MCMC algorithms, such as the Metropolis-Hastings algorithm and the Gibbs sampler. Also, we would like to state that the pop-SAMC algorithm is essentially a dynamic importance sampling algorithm for which the trial distribution, i.e., the working target distribution (3.8), changes from iteration to iteration, and the quantities of interest can be estimated through weighted averaging as in conventional importance sampling (Liang, 2009). That is, Pop-SAMC generates a sequence of importance samples $\{(\boldsymbol{\theta}_{t,1}, e^{\xi_{t,J(\boldsymbol{\theta}_{t,1})}}), \ldots, (\boldsymbol{\theta}_{t,\kappa}, e^{\xi_{t,J(\boldsymbol{\theta}_{t,\kappa})}})\}$, where $J(\boldsymbol{\theta}_{t,i})$ denotes the index of the subregion that $\boldsymbol{\theta}_{t,i}$ belongs to, and $e^{\xi_{t,J(\boldsymbol{\theta}_{t,i})}}$ specifies the importance weight of $\boldsymbol{\theta}_{t,i}$.

OpenMP is an application programming interface (API) for parallel programming on multi-core CPUs which are now available in regular desktops/laptops. It works in a shared memory mode with the fork/join parallelism, and is particularly suitable

for pop-SAMC. To be precise, the population sampling step of pop-SAMC can be carried out in parallel through the pragma *omp parallel* to fork multiple threads with each thread running for an individual Markov chain. After the parallel execution, the threads join back to the master thread, where $\boldsymbol{\xi}_t$ is updated based on the information collected from the multiple threads. Since OpenMP works in a shared memory mode, distributing the updated $\boldsymbol{\xi}_t$ to different threads is avoided. Since the population sampling steps cost the major portion of the CPU, the parallel execution provides a nearly linear speedup for the simulation.

### 3.3.2 Double Parallel Monte Carlo

Based on the subposterior aggregation theory studied in Section 2 and the Pop-SAMC algorithm, we suggest the following double parallel Monte Carlo algorithm for Bayesian analysis of big data.

- (Data Parallel) Divide the dataset into $k$ subsets with each containing about the same sample size.

- (Simulation Parallel) Run Pop-SAMC for each subposterior $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{X}_{[i]})$ separately. Let $\{(\boldsymbol{\theta}_1^{(i)}, w_1^{(i)}), \ldots, (\boldsymbol{\theta}_N^{(i)}, w_N^{(i)})\}$ denote the importance samples generated by Pop-SAMC from $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{X}_{[i]})$ for $i = 1, \ldots, k$. Let $\hat{\boldsymbol{\mu}}^{(i)} = \frac{\sum_{j=1}^N w_j^{(i)} \boldsymbol{\theta}_j^{(i)}}{\sum_{j=1}^N w_j^{(i)}}$ denote the mean of the subposterior $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{x}_{[i]})$.

- (Sample aggregation) Calculate the global mean $\hat{\boldsymbol{\mu}} = \sum_{i=1}^k \hat{\boldsymbol{\mu}}^{(i)}/k$, recenter the importance samples as $\{(\boldsymbol{\theta}_1^{(i)} - \hat{\boldsymbol{\mu}}^{(i)} + \hat{\boldsymbol{\mu}}, w_1^{(i)}), \ldots, (\boldsymbol{\theta}_N^{(i)} - \hat{\boldsymbol{\mu}}^{(i)} + \hat{\boldsymbol{\mu}}, w_N^{(i)})\}$ for $i = 1, \ldots, k$.

Then, for each $i = 1, 2, \ldots, k$, the re-centered samples can be viewed as a batch of importance samples generated from the full data posterior. For any function $h(\boldsymbol{\theta})$ that satisfies $(A4)$, the expectation $\rho = E_\pi h(\boldsymbol{\theta})$ can be naturally estimated by

$\widehat{\rho}_1 = \sum_{i=1}^{k} \widehat{\rho}_1^{(i)}/k$, where $\widehat{\rho}_1^{(i)} = \sum_{j=1}^{N} w_j^{(i)} h(\boldsymbol{\theta}_j^{(i)} - \hat{\boldsymbol{\mu}}^{(i)} + \hat{\boldsymbol{\mu}}) / \sum_{j=1}^{N} w_j^{(i)}$. Alternatively, $\rho$ can be estimated by

$$\widehat{\rho}_2 = \frac{\sum_{i=1}^{k} \sum_{j=1}^{N} w_j^{(i)} h(\boldsymbol{\theta}_j^{(i)} - \hat{\boldsymbol{\mu}}^{(i)} + \hat{\boldsymbol{\mu}})}{\sum_{i=1}^{k} \sum_{j=1}^{N} w_j^{(i)}}.$$

Let $U_i = \sum_{j=1}^{N} w_j^{(i)} h(\boldsymbol{\theta}_j^{(i)} - \hat{\boldsymbol{\mu}}^{(i)} + \hat{\boldsymbol{\mu}})$, $S_i = \sum_{j=1}^{N} w_j^{(i)}$, $S = E(S_i)$, and $V_i = U_i - \rho S_i$. Following from the property of SAMC, the variances of $U_i$ and $V_i$ are both finite. Then the standard error of $\widehat{\rho}_2$ can be calculated as for the ratio estimate (Ripley, 1987). The $V_i$'s can be treated as iid random variables with zero mean and finite variance, and its variance can be estimated by $\hat{\sigma}_V^2 = 1/k \sum_{i=1}^{k} V_i^2$. The law of large numbers implies that $1/\sqrt{k} \sum_{i=1}^{k} V_i$ is asymptotically normal $N(0, \sigma_V^2)$ and that

$$\sqrt{k}(\widehat{\rho}_2 - \rho) = \frac{\frac{1}{\sqrt{k}} \sum_{i=1}^{k} V_i}{\frac{1}{k} \sum_{i=1}^{k} S_i} \to N(0, \sigma^2),$$

where $\sigma^2 = \sigma_V^2 / S^2$, and it can be estimated by $\hat{\sigma}_V^2 / \hat{S}^2$ with $\hat{S} = \sum_{i=1}^{k} S_i / k$.

## 3.4 Simulation Study

### 3.4.1 Logistic Regression

The first example is very simple, whose goal is to show the validity of the proposed subposterior aggregation method. The example is adopted from Srivastava *et al.* (2015). It is for a logistic regression with $n = 10^4$ and the true parameter $\boldsymbol{\theta}^* = (1, -1)^T$. The covariates $Z_1$ and $Z_2$ are drawn from the standard Gaussian distribution. The prior distribution of $\boldsymbol{\theta}$ is $N(0, I_2)$. To follow the notation in Section 3.2, we let $X = (Y, Z_1, Z_2)$.

To implement the proposed double parallel algorithm, we randomly divided the dataset into 10 subsets with each consisting of 1000 samples. Then Pop-SAMC was

run for each subset. Specifically, for each subset, we partitioned the parameter space $\Theta$ according to the energy function $U(\boldsymbol{\theta}) = -\log p_j(\boldsymbol{\theta})$ with an equal bandwidth $\Delta u = 0.5$ into five subregions $E_1 = \{\boldsymbol{\theta} : U(\boldsymbol{\theta}) < u+0.5\}$, $E_2 = \{\boldsymbol{\theta} : u+0.5 \leq U(\boldsymbol{\theta}) < u+1\}$, $E_3 = \{\boldsymbol{\theta} : u+1 \leq U(\boldsymbol{\theta}) < u+1.5\}$ $E_4 = \{\boldsymbol{\theta} : u+1.5 \leq U(\boldsymbol{\theta}) < u+2\}$, and $E_5 = \{\boldsymbol{\theta} : U(\boldsymbol{\theta}) \geq u+2\}$, where $p_j$ denote the subposterior of the $j$th subset, and $u$ was chosen as the smallest value of $U(\boldsymbol{\theta})$ obtained in a preliminary trial. The gain factor $\gamma_t$ was set as $100/\max(100, t)$. The proposal was set as a Gaussian random walk distribution with the covariance matrix $0.2^2 I_2$. The population size was set to $N = 10$ and the number of iterations was set to $T = 10^5$. The first $10^4$ iterations were discarded for the burn-in process, and samples were collected from the remainder of the run at every 5 iterations. In total, we had $1.8 \times 10^5$ importance samples collected at the end of each run.

Figure 3.1 shows the contour plots of the full data posterior $\pi(\boldsymbol{\theta}|\boldsymbol{X})$, each subposterior $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{X}_{[j]})$, and the proposed mixture posterior $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{X})$. The R package KernSmooth was used to generate the corresponding binned kernel density estimates. The plots indicate that each subposterior has a similar shape with the full data posterior, however, most of them have a notably biased center from the true parameter $\boldsymbol{\theta}^*$. By shifting the mean of each subposterior to the global mean, the bias was successfully removed. The mixture posterior $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{X})$ closely matches the full data posterior $\pi(\boldsymbol{\theta}|\boldsymbol{X})$.

### 3.4.2 Linear Regression with Unknown Variance

We use this example to compare the accuracy of the approximations to the full data posterior by the proposed algorithm, WASP(Srivastava *et al.*, 2015) and consensus Monte Carlo(Scott *et al.*, 2016). The example was adopted from Liang *et al.*

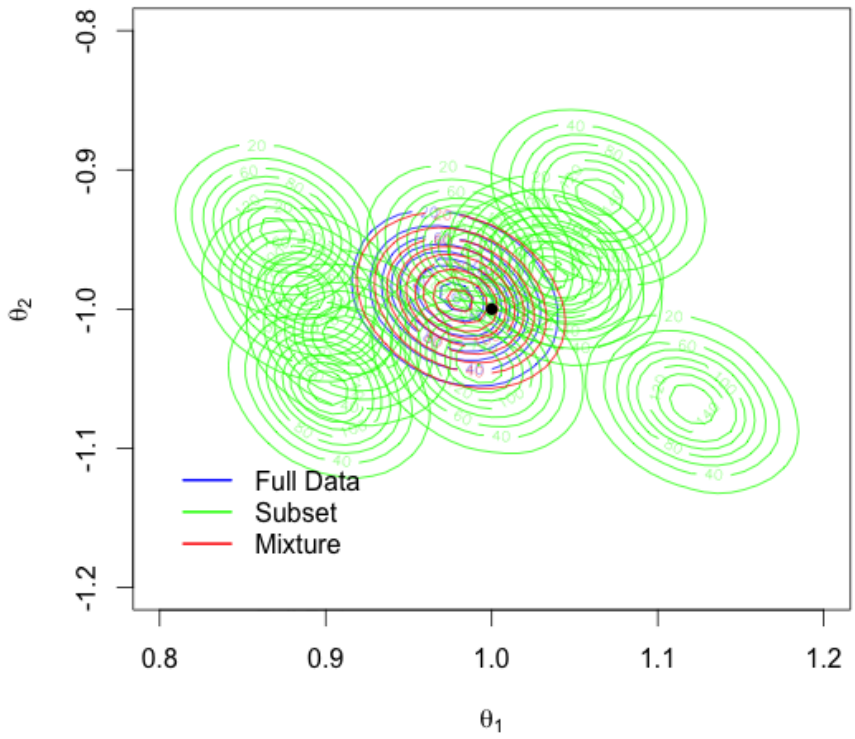


Figure 3.1: Binned kernel posterior density estimates for the parameters of a logistic regression. The true parameter $\boldsymbol{\theta}^* = (1, -1)^T$ (black dot).

(2016), which is about a normal linear regression with unknown variance:

$$y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3} + \epsilon_i, \qquad i = 1, 2, \ldots, n$$

where $(\beta_0, \beta_1, \beta_2, \beta_3) = (2, 0.25, 0.25, 0)$ the true regression coefficients, in addition, $\epsilon_1, \cdots, \epsilon_n$ are i.i.d. normal random errors with mean 0 and variance $\sigma^2 = 0.25$. The covariates $z_1$ and $z_2$ are drawn from standard normal distributions independently. The covariate $z_3 = 0.7z_2 + 0.3e$, where $e$ also follows the standard normal distribution. Under this setting, $z_2$ and $z_3$ are highly correlated with a correlation coefficient of

0.919. We generated $n = 10^4$ samples from this model. For this example, we are to estimate both the regression coefficients and the variance of the random error, i.e., $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2)$. For the regression coefficients, we use non-informative prior $g(\beta_0, \beta_1, \beta_2, \beta_3) \propto 1$; for the variance $\sigma^2$, we use the prior $g(\sigma^2) \propto (\frac{1}{\sigma})^{1/1000}$. To follow the notation in Section 3.2, we set $X_i = (y_i, z_{i1}, z_{i2}, z_{i3})$.

For the double parallel algorithm, we randomly divided the dataset into 10 subsets with each consisting of 1000 samples. Pop-SAMC was run for each subset separately with the same setting as for the previous example except that the energy bandwidth was set to $\Delta u = 2$ and the covariance matrix of the Gaussian random walk proposal distribution was set to $0.01^2 I_5$. For comparison, consensus Monte Carlo and WASP were also applied to this example. For WASP, due to the limitation of memory, we only used 300 posterior samples that were randomly selected from the pool of Metropolis-Hastings samples collected previously. Note that for consensus Monte Carlo, the subset posterior is defined as

$$\frac{\prod_{i=1}^{m} f(X_{ji}|\boldsymbol{\theta}) g^{1/k}(\boldsymbol{\theta})}{\int_{\Theta} \prod_{i=1}^{m} f(X_{ji}|\boldsymbol{\theta}) g^{1/k}(\boldsymbol{\theta}) d\boldsymbol{\theta}}, \tag{3.9}$$

which is slightly different from the subposterior defined in (3.2), the one used in WASP and double parallel.

Figure 3.2 shows the QQ-plots for each of the five parameters of the model and for each of the methods, double parallel, consensus Monte Carlo and WASP, versus the full data posterior simulation. The QQ plots indicate that double parallel and consensus Monte Carlo can provide more accurate approximations to the full data posterior than WASP. Regarding efficiency of the three algorithms, we compared the rough number of effective samples produced by them with the same CPU time. Within a given CPU time, the double parallel algorithm produced $1.8 \times 10^6$

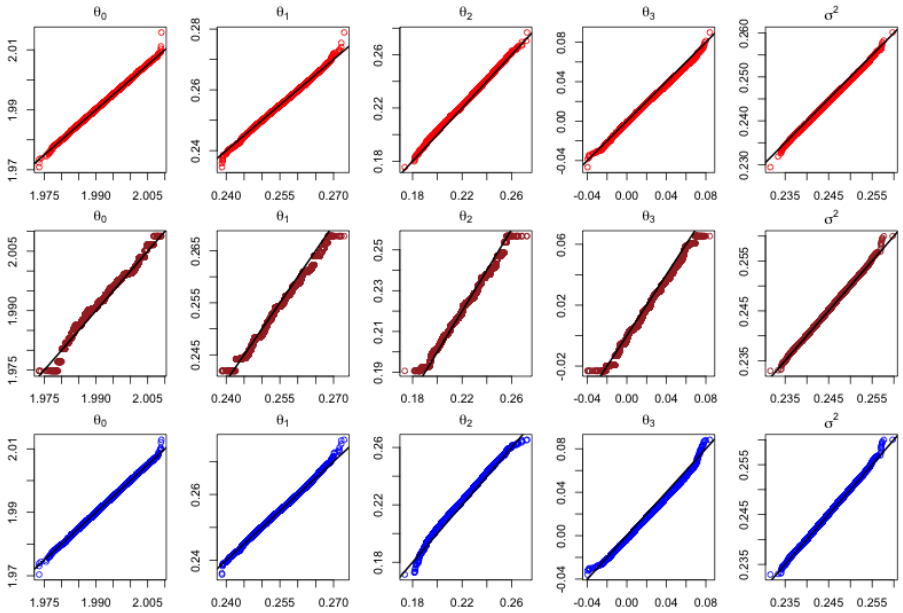


Figure 3.2: QQ-plots for the normal regression example. The top, middle and bottom panels are for the double parallel, WASP and consensus Monte Carlo, respectively.

importance samples ($1.8 \times 10^5$ importance samples were collected for each of the 10 subsets). However, consensus Monte Carlo produced only $1.8 \times 10^5$ samples, for which the samples produced by different chains (each for a different subset) are averaged to get the final samples. For WASP, the samples produced by different chains do not need to be averaged, but need to be weighted through linear programming in estimating their Wasserstein barycenter. Again, the importance weighting procedure will significantly reduce its effective sample size.

### 3.5 A Big Data Example

The goal of this example is to show how efficient the double parallel algorithm can be compared to the traditional single chain MCMC algorithm for a big data problem. For this purpose, we applied the double parallel algorithm to the MiniBooNE particle identification dataset, which is available at the UCI machine learning repository.

This dataset records 130,064 events (observations), including 36,499 signal events and 93,565 background events. Each observation consists of the event type (signal event or background event) and 50 associated particle variables. The task of the problem is to explore the relationship between the event type and the associated particle variables. The more detailed description for the dataset and its physical background can be found in Roe *et al.* (2005).

The problem can be naturally treated using a logistic regression, where the event type is used as the response variable and the 50 associated particle variables are used as the predictors. To identify the important variables that are associated with the event type, we let the regression coefficients be subject to a heavy-tail distribution, $t(3)$, which belongs to the class of local shrinkage priors but is more moderate in shrinking large regression coefficients than the Lasso prior(Tibshirani, 1996). Other local shrinkage priors, such as the horseshoe prior(Carvalho *et al.*, 2010), can also be used here without affecting on the efficiency of the proposed algorithm. In data preprocessing, we first removed 468 samples with missing observations and then randomly divide the remaining samples into 10 subsets of nearly the same sample size. For each subset, Pop-SAMC was run with the population size $\kappa = 20$. The sample space was partitioned with an energy bandwidth $\Delta u = 1$ and the subregions determined through a preliminary run. The gain factor was set as $\gamma_t = \min(1, (t/1000)^{-0.6})$. The algorithm was run for $1.1 \times 10^5$ iterations, where the first $10^4$ iterations were for the burn-in process and the samples collected from the remaining $10^5$ iterations were used for inference. At each iteration, 20 parameters was randomly selected to be updated along a random direction with a step size of 0.1. The acceptance rate was around 0.16, which indicates the effectiveness of the simulation. On a high-end Dell Precision T7610 Workstation with 24 cores, one run of Pop-SAMC costs about 9 minutes (wall clock time) or 166 minutes total CPU

time.

For comparison, the single chain Metropolis-Hastings (MH) algorithm was directly applied to simulate from the full data posterior. The algorithm was run for $2.2 \times 10^6$ ($= 20 \times 1.1 \times 10^5$) iterations correspondingly, where the first $2.0 \times 10^5$ were discarded for the burn-in process and the samples collected from the remaining iterations were used for inference. The MH algorithm used the same proposal as the Pop-SAMC and the resulting acceptance rate was also about 0.16. On the same computer, one run of the MH algorithm costs about 1,373 minutes (wall clock time) or 1,371 minutes CPU time. In wall clock time, the computational cost by the double parallel algorithm is only 0.67% of that by the single chain MH algorithm!

Table 3.1 shows the computational results produced by the two methods. For each method, we reported only the ten most significant variables, including their posterior mean and standard deviation. Here the significance of each variable was measured according to the ratio of its posterior mean and standard deviation. The results from the two algorithms are surprisingly consistent: All the variables have about the same posterior mean and standard deviation. The top 10 significant variables are exactly the same, even with the same order! This again indicates the validity of the double parallel algorithm.

Table 3.1: Comparison of computational time (wall clock time) and parameter estimation for the MiniBooNE particle data set.

| Double Parallel($k = 10$, $\kappa = 20$) | | MH Algorithm | |
|---|---|---|---|
| wall clock time | | | |
| 9.2 minutes | | 1,373 minutes | |
| Top 10 significant variables | | | |
| var13 | -1.2273 (0.0153) | var13 | -1.2140 (0.0158) |
| var1 | -1.3667 (0.0338) | var1 | -1.3529 (0.0328) |
| var16 | 3.4174 (0.0861) | var16 | 3.3730 (0.0840) |
| var4 | -0.8825 (0.0234) | var4 | -0.8788 (0.0254) |
| var32 | 1.0733 (0.0329) | var32 | 1.0590 (0.0309) |
| var17 | -2.1592 (0.0709) | var17 | -2.1169 (0.0672) |
| var6 | 0.3862 (0.0134) | var6 | 0.3825 (0.0139) |
| var12 | -0.7624 (0.0281) | var12 | -0.7503 (0.0278) |
| var34 | -0.9335 (0.0344) | var34 | -0.9239 (0.0355) |
| var25 | 0.4082 (0.0206) | var25 | 0.4004 (0.0206) |

# 4. AVERAGE BAYESIAN INFORMATION CRITERION AND ITS APPLICATION TO HIGH DIMENSIONAL GENERALIZED LINEAR MODEL

## 4.1 Introduction

Model selection is an important part of any statistical analysis. For example, in polynomial regression one has to determine the degree of the polynomial; in multivariate regression one needs to select the covariates included into the model; in stationary time series one should choose the order for their ARMA model. Due to the importance and popularity of this topic, numerous model selection approaches have been proposed in the past, such as Mallows's $C_p$ (Mallows, 1973), AIC (Akaike, 1974), cross validation (Stone, 1974), BIC (Schwarz, 1978), generalized cross validation (Wahba, 1979), RIC (Foster and George, 1994), Bayes model avaraging (Raftery *et al.*, 1997) and DIC (Spiegelhalter *et al.*, 2002). Among these methods, some can be classfied into "information criterion" family. These criteria try to strike a balance between the model's fitting performance, usually measured by its maximized log-likelihood, and its complexity, usually measured by a penalty term involving the size of the model.

Akaike's information criterion (Akaike, 1974) is the first and most famous member of this family, which is defined as

$$\mathrm{AIC}(s) = -2l(D_n|\widehat{\boldsymbol{\beta}}_s) + 2|s|$$

where $s$ denotes a specific model, $|s|$ denotes the size of the model $s$, $l$ denotes log-likelihood function, $D_n$ denotes observed data, $\widehat{\boldsymbol{\beta}}_s$ denotes the MLE of the parameters

$\boldsymbol{\beta}$ in model $s$. AIC has a pretty good predictive performance, but it is not a consistent criterion. As the number of observations $n$ grows infinitely large, AIC is not guaranteed to choose the true data generating model. Instead, it often tends to select more complex models that overfit the data. To overcome this problem, Bayesian information criterion was proposed(Schwarz, 1978). This criterion is derived from the bayesian perspective and imposes heavier penalty on the model size, that is

$$\mathrm{BIC}(s) = -2l(D_n|\widehat{\boldsymbol{\beta}}_s) + |s|\log(n)$$

Under specific conditions, BIC has been shown to be a consistent criterion and can accurately select the smallest true model when $n$ is large enough. See, for instance, Nishii (1984); Haughton (1988).

Another interesting and important criterion is Deviance information criterion (Spiegelhalter *et al.*, 2002), which is defined as

$$\mathrm{DIC}(s) = -2\bar{l}(D_n|\boldsymbol{\beta}_s) + (-2)[\bar{l}(D_n|\boldsymbol{\beta}_s) - l(D_n|\bar{\boldsymbol{\beta}}_s)]$$

where $\bar{l}(D_n|\boldsymbol{\beta}_s) = E_{\boldsymbol{\beta}_s|D_n,s}l(D_n|\boldsymbol{\beta}_s)$ denotes the posterior mean of the loglikelihood function for model $s$ and $\bar{\boldsymbol{\beta}}_s = E(\boldsymbol{\beta}_s|D_n, s)$ denotes the posterior mean of $\boldsymbol{\beta}_s$ for model $s$. When sample size $n$ is large enough, we have $[\bar{l}(D_n|\boldsymbol{\beta}_s) - l(D_n|\bar{\boldsymbol{\beta}}_s)] \approx -|s|/2$ and $l(D_n|\bar{\boldsymbol{\beta}}_s) \approx l(D_n|\widehat{\boldsymbol{\beta}}_s)$, thereby DIC can be viewed as an approximation to AIC with posterior samples. Unfortunately, due to its similarity with AIC, DIC is not a consistent criterion, either. Inspired by the idea of appximating maximum log-likelihood function by posterior samples, we developed a new criterion, Average

Bayesian Information Criterion,

$$\text{ABIC}(s) = -2\bar{l}(D_n|\boldsymbol{\beta}_s) - |s| + |s|\log(n)$$

This time, instead of approximating AIC, we approximate BIC using posterior samples. Since BIC is a consistent criterion, it is reasonable that ABIC is also consistent. In addition, for $n$ sufficiently large, the term $-|s|$ is ignorable, and we can rewrite ABIC as

$$\text{ABIC}(s) = -2\bar{l}(D_n|\boldsymbol{\beta}_s) + |s|\log(n)$$

We want to point out that this name should be better interpreted as "a BIC-like information criterion made from posterior Average of log-likelihoods ", rather than "Average of BIC".

Recently, high-dimensional data has become very popular in many areas of modern scientific research, such as genomics, microarrays, proteomics and brain images. For example, in genowide association studies between genotypes and phenotypes, millions of SNPs are potential covariates; in disease classfication using microarray or proteomics data, thousands of expression profiles are potential predictors. Moreover, when interaction are considered, the dimensionality will grow more quickly. These massive amounts of high dimensional data bring not only opportunities but also lots of challenges to statistical inference (Fan and Li, 2007; Johnstone and Titterington, 2009). Particulary, in this high-dimensional setting, where the sample size $n$ is smaller than the dimension of parameters $p$, many traditional model selection methods fail to maintain their good property, some even become non-implementable. Therefore, there is increasing need to develop new techniques to select models in this high-dimensional situation. In fact, some have already been proposed, to name a

few, Lasso(Tibshirani, 1996), SCAD(Fan and Li, 2001), elastic net (Zou and Hastie, 2005), Dantzig selector (Candes and Tao, 2007), MCP (Zhang, 2010) and rLasso (Song and Liang, 2015),

As with many of its fellows, BIC is no longer a consistent criterion under the high-dimensional setting. Instead, it is usually too literal, tending to select a model with spurious covariates. To deal with this problem, Chen and Chen (2008) developed extended Bayesian information criterion (EBIC), which is defined as

$$\text{EBIC}(s) = -2l(D_n|\widehat{\boldsymbol{\beta}}_s) + |s|\log(n) + 2\gamma|s|\log(p)$$

where $p$ denotes the total number of candidate parameters. Under some sparsity and regularity conditions, EBIC has been proved to be a consistent information criterion for both linear model and generalized linear model. (Chen and Chen, 2008, 2012; Chen and Luo, 2013). Correspondingly, ABIC can also be modified to Average Extended Bayesian Information Criterion (AEBIC)

$$\text{AEBIC}(s) = -2\bar{l}(D_n|\boldsymbol{\beta}_s) + |s|\log(n) + 2\gamma|s|\log(p)$$

to conduct model selection under the high-dimensional setting. Although AEBIC should be applicable to a broad class of models, in this chapter we limit ourselves to the generalized linear model. Under some sparsity and regularity conditions, the consistency property of AEBIC is also established.

The remainder of this chapter is organized as follows. Section 4.2 describes an informal derivation of AEBIC and a detailed algorithm for conducting model selection based on this information criterion. Section 4.3 establishes the consistency property of AEBIC under some assumptions. Section 4.4 and section 4.5 evaluates the finite

sample performance of ABIC and AEBIC through some simulation studies and a real data example, respectively.

### 4.2 Average Extended Bayesian Information Criterion

Let $D_n = \{(y^{(i)}, \mathbf{x}^{(i)}) : i = 1, \ldots, n\}$ denote a dataset of n observations, where the explanatory variable $\mathbf{x}$ is a $p_n$-dimensional random vector. In high dimensional setting, $p_n$ can increase with the sample size n. Assume the conditional distribution of $y$ given $\mathbf{x}$ follows a parametric generalized linear model (GLM) (McCullagh and Nelder, 1989) with the following form.

$$f(y|\mathbf{x}, \boldsymbol{\beta}) = \exp(\theta y - b(\theta) + c(y))$$

where $b(\cdot)$ is continuously differentiable and $\theta$ is the natural parameter, which relates $y$ to the predictors via a linear function

$$\theta = \mathbf{x}^T \boldsymbol{\beta} = \beta_1 x_1 + \cdots + \beta_{p_n} x_{p_n}$$

Here the intercept term has been treated as a special predictor included in $\mathbf{x}$. In this model, the mean function $u = E(y|\mathbf{x}) = b'(\theta)$. This class of GLMs includes poisson regression, logistic regression and linear regression (with known variance).

In reality, the true parameter $\boldsymbol{\beta}^*$ may contain lots of zero components. If we let $s$ be a subset of $\{1, \ldots, p_n\}$ and S be a set consisting of all such subsets, then each $s$ can specify an individual model and $S$ is just the model space. Accordingly, we use $s^*$ to denote the true model, that is, the subset consisting of nonzero component indexes of $\boldsymbol{\beta}^*$. The objective of model selection is to correctly find $s^*$ from all possible $s \in S$, based on the observed data.

In Bayesian perspective, we prefer to choose the model with maximum posterior

probability. The likelihood function(ignoring the marginal density of $\mathbf{x}$) of observed data for model $s$ is

$$L(D_n|\boldsymbol{\beta}_s) = \prod_{i=1}^{n} f(y^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\beta}_s)$$

where $\boldsymbol{\beta}_s$ denotes the components in $\boldsymbol{\beta}$ corresponding to model $s$. If we further let $\pi(s)$ denote the prior probability of model $s$ and $\pi(\boldsymbol{\beta}(s))$ denote the prior probability of parameters in model $s$, then the posterior probability of model $s$, $P(s|D_n)$ is

$$P(s|D_n) = \frac{m(D_n|s)\pi(s)}{\sum_{s\in S} m(D_n|s)\pi(s)}$$

where

$$m(D_n|s) = \int L(D_n|\boldsymbol{\beta}(s))\pi(\boldsymbol{\beta}(s))d\boldsymbol{\beta}(s)$$

By choosing some appropriate priors, it can be shown that $P(s^*|D_n)$ converges to 1 in probability as $n$ goes to infinity (Liang *et al.*, 2013), which is the so-called global model consistency in Bayesian variable selection (Johnson and Rossell, 2012). This further implies that $P(\arg\max_s P(s|D_n) = s^*)$ converges to 1 in probability as $n$ goes to infinity.

But in most cases, it's impossible to calculate $P(s|D_n)$ exactly, so we need to approximate it. One way is to use MCMC samples to approximate it. Another way is to use Laplace approximation to deal with it, that is also how EBIC(Chen and Chen, 2012) is derived:

To be more specific, since

$$\arg\max_s P(s|D_n) = \arg\max_s m(D_n|s)\pi(s) = \arg\max_s \log\{m(D_n|s)\pi(s)\}$$

From now on, we only care about $\log\{m(D_n|s)\pi(s)\}$,

$$\log\{m(D_n|s)\pi(s)\} = \log \int L(D_n|\boldsymbol{\beta}_s)\pi(\boldsymbol{\beta}_s)d\boldsymbol{\beta}_s + \log \pi(s)$$

To deal with the integral term, we first need to expand $\log L(D_n|\boldsymbol{\beta}_s)$ at $\widehat{\boldsymbol{\beta}}_s$, which is the maximum likelihood estimator of $\boldsymbol{\beta}_s$ given model $s$.

$$\begin{aligned}
\log L(D_n|\boldsymbol{\beta}_s) &\approx \log L(D_n|\widehat{\boldsymbol{\beta}}_s) + \frac{1}{2}(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)'\left[\frac{\partial^2 \log L(D_n|\widehat{\boldsymbol{\beta}}_s)}{\partial\boldsymbol{\beta}_s\partial\boldsymbol{\beta}_s^T}\right](\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s) \\
&= \log L(D_n|\widehat{\boldsymbol{\beta}}_s) - \frac{1}{2}(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)'\left[n\widehat{I}(D_n, \widehat{\boldsymbol{\beta}}_s)\right](\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)
\end{aligned}$$

where

$$\widehat{I}(D_n, \widehat{\boldsymbol{\beta}}_s) = -\frac{1}{n}\frac{\partial^2 \log L(D_n|\widehat{\boldsymbol{\beta}}_s)}{\partial\boldsymbol{\beta}_s\partial\boldsymbol{\beta}_s^T}$$

is the averaged observed information matrix. To step it further,

$$L(D_n|\boldsymbol{\beta}_s) \approx L(D_n|\widehat{\boldsymbol{\beta}}_s)\exp\{-\frac{1}{2}(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)'\left[n\widehat{I}(D_n, \widehat{\boldsymbol{\beta}}_s)\right](\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)\}$$

Now we have the following approximation for the integral term

$$\begin{aligned}
&\int L(D_n|\boldsymbol{\beta}_s)\pi(\boldsymbol{\beta}_s)d\boldsymbol{\beta}_s \\
\approx\ & L(D_n|\widehat{\boldsymbol{\beta}}_s)\int \exp\{-\frac{1}{2}(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)'\left[n\widehat{I}(D_n, \widehat{\boldsymbol{\beta}}_s)\right](\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)\}\pi(\boldsymbol{\beta}_s)d\boldsymbol{\beta}_s \\
\approx\ & L(D_n|\widehat{\boldsymbol{\beta}}_s)\pi(\widehat{\boldsymbol{\beta}}_s)\int \exp\{-\frac{1}{2}(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)'\left[n\widehat{I}(D_n, \widehat{\boldsymbol{\beta}}_s)\right](\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)\}d\boldsymbol{\beta}_s \\
=\ & L(D_n|\widehat{\boldsymbol{\beta}}_s)\pi(\widehat{\boldsymbol{\beta}}_s)(2\pi)^{\frac{|s|}{2}}|n\widehat{I}(D_n, \widehat{\boldsymbol{\beta}}_s)|^{-\frac{1}{2}} \\
=\ & L(D_n|\widehat{\boldsymbol{\beta}}_s)\pi(\widehat{\boldsymbol{\beta}}_s)(\frac{2\pi}{n})^{\frac{|s|}{2}}|\widehat{I}(D_n, \widehat{\boldsymbol{\beta}}_s)|^{-\frac{1}{2}}
\end{aligned}$$

where $|s|$ is the size of model $s$. The second approximation is valid provided that the

prior $\pi(\boldsymbol{\beta}_s)$ is "flat" over the neighborhood of $\widehat{\boldsymbol{\beta}}_s$ and $L(D_n|\boldsymbol{\beta}_s)$ is ignorable outside the neighborhood of $\widehat{\boldsymbol{\beta}}_s$.

Finally,

$$\begin{aligned}
\log\{m(D_n|s)\pi(s)\} &= \log \int L(D_n|\boldsymbol{\beta}_s)\pi(\boldsymbol{\beta}_s)d\boldsymbol{\beta}_s + \log \pi(s) \\
&\approx \log L(D_n|\widehat{\boldsymbol{\beta}}_s) + \log \pi(\widehat{\boldsymbol{\beta}}_s) + \frac{|s|}{2}\log(2\pi) - \frac{|s|}{2}\log(n) \\
&\quad -\frac{1}{2}\log|\widehat{I}(D_n, \widehat{\boldsymbol{\beta}}_s)| + \log \pi(s) \\
&\approx \log L(D_n|\widehat{\boldsymbol{\beta}}_s) - \frac{|s|}{2}\log(n) + \log \pi(s)
\end{aligned}$$

The second approximation is justified by the fact that for n sufficiently large, $log(2\pi)$ is comparably ignorable than $log(n)$,$|\widehat{I}(D_n, \widehat{\boldsymbol{\beta}}_s)|$ converges to a constant and $\log \pi(\widehat{\boldsymbol{\beta}}_s)$ can be usually controlled as $O(1)$.

For $\pi(s)$, if we adopt the setting in Liang *et al.* (2013)

$$\pi(s) = \lambda_n^{|s|}(1-\lambda_n)^{p_n-|s|}$$

where $\lambda_n$ denotes the probability of each individual variable to be selected for model $s$ and is taken a value of the form

$$\lambda_n = \frac{1}{1+p_n^{\gamma}\sqrt{2\pi}}$$

for some parameter $\gamma > 0$; or adopt the setting in Chen and Chen (2008), that is

$$\pi(s) \propto \binom{p_n}{|s|}^{-\gamma},$$

then for $|s|$ far smaller than $P_n$, it is easy to verify

$$\log \pi(s) = -\gamma |s| \log(p_n) + O(1)$$

Therefore, we can further have

$$\log\{m(D_n|s)\pi(s)\} \approx \log L(D_n|\widehat{\boldsymbol{\beta}}_s) - \frac{|s|}{2}\log(n) - \gamma|s|\log(p_n)$$

which exactly equals $-\frac{1}{2}$EBIC

In this chapter, we propose a new method to approximate $\log\{m(D_n|s)\pi(s)\}$. To be more specific, starting from the second method, we can further approximate $\log L(D_n|\widehat{\boldsymbol{\beta}}_s)$ by $E_{\{\boldsymbol{\beta}_s|D_n,s\}} \log L(D_n|\boldsymbol{\beta}_s) + \frac{|s|}{2}$, where $\{\boldsymbol{\beta}_s|D_n, s\}$ denote the posterior distribution of $\boldsymbol{\beta}_s$ given data $D_n$ and model $s$. This approximation has also been discussed in Spiegelhalter *et al.* (2002). To verify this, we notice that

$$\begin{aligned} P(\boldsymbol{\beta}_s|D_n, s) &= \frac{L(D_n|\boldsymbol{\beta}_s)\pi(\boldsymbol{\beta}_s)}{m(D_n|s)} \propto L(D_n|\boldsymbol{\beta}_s)\pi(\boldsymbol{\beta}_s) \\ &\approx L(D_n|\widehat{\boldsymbol{\beta}}_s)\pi(\widehat{\boldsymbol{\beta}}_s)\exp\{-\frac{1}{2}(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)'\left[n\widehat{I}(D_n, \widehat{\boldsymbol{\beta}}_s)\right](\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)\} \end{aligned}$$

for n sufficiently large, at the neighborhood of $\widehat{\boldsymbol{\beta}}_s$, where $L(D_n|\boldsymbol{\beta}_s)$ is dominant. Therefore, $\{\boldsymbol{\beta}_s|D_n, s\}$ is asymptotic normal with mean $\widehat{\boldsymbol{\beta}}_s$ and covariance matrix

$\left[n\widehat{I}(D_n, \widehat{\boldsymbol{\beta}}_s)\right]^{-1}$. Now we have

$$\begin{aligned}
& E_{\{\boldsymbol{\beta}_s|D_n,s\}} \log L(D_n|\boldsymbol{\beta}_s) \\
\approx\ & E_{\{\boldsymbol{\beta}_s|D_n,s\}}\{\log L(D_n|\widehat{\boldsymbol{\beta}}_s) - \frac{1}{2}(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)'\left[n\widehat{I}(D_n, \widehat{\boldsymbol{\beta}}_s)\right](\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)\} \\
=\ & \log L(D_n|\widehat{\boldsymbol{\beta}}_s) - \frac{1}{2}E_{\{\boldsymbol{\beta}_s|D_n,s\}}(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)'\left[n\widehat{I}(D_n, \widehat{\boldsymbol{\beta}}_s)\right](\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s) \\
=\ & \log L(D_n|\widehat{\boldsymbol{\beta}}_s) - \frac{1}{2}|s|
\end{aligned}$$

which verified our approximation. Thus, we can approximate $\log\{m(D_n|s)\pi(s)\}$ by the following

$$\begin{aligned}
\log\{m(D_n|s)\pi(s)\} &\approx \log L(D_n|\widehat{\boldsymbol{\beta}}_s) - \frac{|s|}{2}\log(n) - \gamma|s|\log(p_n) \\
&\approx E_{\{\boldsymbol{\beta}_s|D_n,s\}} \log L(D_n|\boldsymbol{\beta}_s) + \frac{|s|}{2} - \frac{|s|}{2}\log(n) - \gamma|s|\log(p_n) \\
&\approx E_{\{\boldsymbol{\beta}_s|D_n,s\}} \log L(D_n|\boldsymbol{\beta}_s) - \frac{|s|}{2}\log(n) - \gamma|s|\log(p_n)
\end{aligned}$$

In the large sample setting, $\frac{|s|}{2}$ is negligible because it is of lower order of $\frac{|s|}{2}\log(n)$.

As with EBIC, we define AEBIC as -2 times the previous expression

$$AEBIC(s) = -2E_{\{\boldsymbol{\beta}_s|D_n,s\}} \log L(D_n|\boldsymbol{\beta}_s) + |s|\log(n) + 2\gamma|s|\log(p_n)$$

so as to be on the deviance scale. In many cases, $E_{\{\boldsymbol{\beta}_s|D_n,s\}} \log L(D_n|\boldsymbol{\beta}_s)$ doesn't have closed-form expression, so we should use MCMC samples to approximate it. The detailed algorithm for the model selection based on AEBIC is described below

1. Use MCMC to generate T samples $\{\boldsymbol{\beta}^{(t)}, s^{(t)}\},\ t = 1, \ldots, T$

2. Use set $S'$ to include all models appearing in $s^{(t)}, t = 1, \ldots, T$

3. For each $s \in S'$, calculate

$$AEBIC(s) = \frac{-2}{\#\{t : s^{(t)} = s\}} \sum_{\{t:s^{(t)}=s\}} \log L(D_n|\boldsymbol{\beta}^{(t)}) + |s| \log(n) + 2\gamma|s| \log(p_n)$$

4. Select $\widehat{s}$, which has the smallest AEBIC among all $s \in S'$, as our estimator for $s^*$

Remark:

- Our method uses both Laplace approximation and MCMC samples, therefore can be viewed as a combination of the two popular methods mentioned before.
- Compared to the first method, which directly uses MCMC samples to approximate $P(s|D_n)$, our method is generally more accurate.
- Compared to the BIC-like methods, our method has at least two merits: First, it need not to calculate MLE, which is sometimes very difficult to obtain. Second, in BIC-like methods,we should select a sequence of candidate models in advance, and this selection is generally completed by other procedures, such as stepwise regression or LASSO with different tuning parameters. However, in our method, this 'selection' is completed automatically, because we only calculate the AEBIC for models appearing in the MCMC samples.

### 4.3 Consistency

In Section 4.2, we give an informal derivation and explanation of our method. In this section, we'll show the statistical properties of our method in a more rigorous way.

At first we present some assumptions (we treat the covariates as fixed in this paper)

(A1) $p_n = O(n^\kappa)$ for some positive constant $\kappa$.

(A2) $|x_j| \leq 1, \text{for } j = 1, \ldots, p_n$

(A3) $|s^*| \leq q$ for a fixed integer $q \in \mathbb{N}$

(A4) $a_1 \leq \lambda_{\min}(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_s^{(i)}(\mathbf{x}_s^{(i)})^T) \leq \lambda_{\max}(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_s^{(i)}(\mathbf{x}_s^{(i)})^T) \leq a_2$ for all model $s$ with $|s| \leq 2q$, where $\mathbf{x}_s$ denotes the sub-vector of $\mathbf{x}$, with respect to model $s$. $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the mininum and maximum eigenvalue, respectively.

(A5) $||\boldsymbol{\beta}^*||_2 \leq M$ for a positive number $M$

(A6) There exist a constant $c_0$ such that $\min_{j \in s^*} |\beta_j^*| \geq c_0 n^{-1/4}$, where $|\beta_j^*|$ denotes the corresponding element in the vector $\boldsymbol{\beta}^*$

(A1) allows $p_n$ to grow polynomially with sample size $n$. (A2) bounds the absolute value of the covariates by 1, this assumption can be easily extended to the case where all $|x_j|$ are bounded above by a larger constant. (A3) bounds the size of the true model, although the true model can be changed with $n$. (A4) bounds the eigenvalue of the second moment matrices uniformly for all models $s$ with $|s| \leq 2q$. (A5) gives an upper bound for the norm of the true parameter. (A6) requires the signal can not be too small to be detected.

Since we have already assumed the true model size $|s^*|$ is smaller than $q$, we can further adopt a modified prior for the models

$$\pi(s) \propto \lambda_n^{|s|}(1-\lambda_n)^{p_n - |s|} I[|s| \leq q]$$

It is easy to see this modification won't affect the validity of the approximation of $\pi(s)$ in Section 4.2

Conditional on $s$, we set the priors for the parameters as follows

$$\pi(\boldsymbol{\beta}_s) = N(0, V_s) = N(0, \sigma_s^2 I)$$

where $\sigma_s^2 = \frac{1}{2\pi} e^{C_0/|s|}$ for some positive constants $C_0$.

Under these assumptions and prior settings, we now present two theorems. The first theorem provides a theoretical validation for approximating $\log L(D_n|\widehat{\boldsymbol{\beta}}_s)$ by $E_{\{\boldsymbol{\beta}_s|D_n,s\}} \log L(D_n|\boldsymbol{\beta}_s)$. The second theorem finally leads to the consistency of AEBIC. The proofs of these two theorems can be found in the Appendix

**Theorem 4.1.** *Assume conditions (A1)-(A5) are satisfied and $\pi(\boldsymbol{\beta}_s)$ is specified as above. Then for sufficiently large $n$, with probability at least $1 - n^{-\eta}$ ($\eta$ can be any positive number), it holds that for all models $s$ with $|s| \leq q$,*

$$E_{\{\boldsymbol{\beta}_s|D_n,s\}} \log L(D_n|\boldsymbol{\beta}_s) - \log L(D_n|\widehat{\boldsymbol{\beta}}_s) = -\frac{|s|}{2} + o(1)$$

**Theorem 4.2.** *Assume conditions (A1)-(A6) are satisfied, $\pi(\boldsymbol{\beta}_s)$ and $\pi(s)$ are specified as above. Moreover, we set $\gamma > 1 - \frac{1-2\eta}{2\kappa}$ for $0 < \eta < \frac{1}{2}$. Then for sufficiently large $n$, with probability at least $1 - n^{-\eta}$, it holds that for all models $s$ with $|s| \leq q$ and $s \neq s^*$*

$$AEBIC(s) > AEBIC(s^*)$$

### 4.4 Simulation

In this section, we present some simulation studies in both low-dimensional and high-dimensional cases. Although the previous sections are talking about high-

dimensional cases, it is obvious that this algorithm remains consistent under the low-dimensional setting.

### 4.4.1 Low Dimensional Logistic Regression

Let's consider a simple logistic regression with only three active features

$$\text{logit } P(y = 1|\mathbf{x}) = 2x_1 + x_2 + 3x_3$$

Here we include 50 candidate features, $x_1, \cdots, x_{50}$, generated from standard normal distributions and sample size $n$ is set as 500.

We first need to specify the priors. For $\pi(s)$, we just let $\gamma = 0$ and ignored the upper bound $q$ since it is a low dimensional case. For $\pi(\boldsymbol{\beta}_s)$, we let $C_0 = 10$. In order to generate posterior samples of $(s, \boldsymbol{\beta}_s)$, we implemented Metropolis Hasting algorithm, where the MH moves include three types: birth, death and parameter updates. In the birth step, we randomly selected a $\beta_i$ outside current model $s$, and added it into $s$ to produce a larger model $s'$. The value of the new member $\beta_i$ was generated from the distribution $N(0, 10^2)$. In the death step, we randomly selected a $\beta_i$ in our current model $s$, and deleted it from the model to obtain a smaller model $s'$. In the parameter update step, we kept the current model $s$, and randomly selected one parameter $\beta_i \in s$ and updated its value by $f(\beta_i^{\text{new}}|\beta_i) = N(0, 0.5^2)$. The probability of the three types was set as $(0.25, 0.25, 0.5)$. In order to accelerate the convergence rates, more advanced algorithms should be implemented. We generated $10^6$ samples in total. The first $10^5$ were discarded as "burn-in" samples. For the rest, we only kept every 10th samples, to reduce the autocorrelation effect.

After we collected useful samples of $(s, \boldsymbol{\beta}_s)$, we implemented the procedures described in Section 4.2 to calculate ABIC for different models and chose the model with smallest ABIC. Notice in the low dimensional case, ABIC doesn't include the

term $\gamma \log(P)$. With these posterior samples, we can also calculate AIC,BIC and DIC for different models. In practice, it's not necessary to calculate the four criteria for each occurring model and here we only considered models having occurred more than 100 times.

We repeated this simulation 100 times and summarized the results in the following table. We used three metrics to evaluate the performance of each method.

1 Mean and standard deviation of the selected model $\widehat{s}$ s' sizes across 100 datasets

2 Positive Selection Rate: calculated by $\frac{\sum_{i=1}^{100} |s^* \cap \widehat{s}|}{|s^*| \cdot 100}$

3 True Discovery Rate: calculated by $\frac{\sum_{i=1}^{100} |s^* \cap \widehat{s}|}{\sum_{i=1}^{100} |\widehat{s}|}$

Table 4.1: Simulation results of AIC,BIC,DIC and ABIC for the low dimensional logistic regression

| | AIC | BIC | DIC | ABIC |
|---|---|---|---|---|
| Mean of Size (SD of size) | 8.86(1.99) | 3.72(0.96) | 9.04(2.15) | 3.33(0.57) |
| Positive Selection Rate | 1.000 | 1.000 | 1.000 | 1.000 |
| True Discovery Rate | 0.339 | 0.806 | 0.332 | 0.901 |

From table 4.1, it is obvious that all methods can choose a model including $x_1, x_2$ and $x_3$. But AIC and DIC often choose a larger model, while BIC and ABIC seldomly adds redundant variables into its model. This observation is in good agreement with the theoretical result, that is, BIC and ABIC are consistent model selection criteria, while AIC and DIC are not.

### 4.4.2 Low Dimensional Linear Regression

Next consider a linear regression with five active features

$$y = x_1 + 2.2x_2 - 1.6x_3 + 2x_4 - 1.4x_5 + \epsilon$$

where $\epsilon$ follows normal distribution with known variance 1. We still included 50 candidate features, $x_1, \cdots, x_{50}$, independently generated from $N(0, 1)$ and set sample size $n$ as 500. Similar procedure were implemented to generate posterior samples of $(s, \boldsymbol{\beta}_s)$. $10^6$ samples were generated in total and the first $10^5$ were discarded as "burn-in" samples. For the rest, we only keft every 10th sample, as before. After collecting useful samples, we calculated AIC, BIC, DIC, ABIC for each model which have occurred more than 100 times , and then selected model based on these four criteria, separately. We repeated this simulation 100 times and summarized the results in the following table. The same three metrics were used again to evaluate the performance of each criteria.

Table 4.2: Simulation results of AIC,BIC,DIC and ABIC for the low dimensional linear regression

| | AIC | BIC | DIC | ABIC |
|---|---|---|---|---|
| Mean of Size (SD of size) | 8.70(1.15) | 5.59(0.82) | 9.05(1.08) | 5.35(0.74) |
| Positive Selection Rate | 1.000 | 1.000 | 1.000 | 1.000 |
| True Discovery Rate | 0.575 | 0.894 | 0.552 | 0.934 |

From table 4.2 we can obtain similar conclusions as those of the previous example. All methods can include the true predictors $x_1$ to $x_5$. However, AIC and DIC often lead to a larger model, while BIC and AveBIC can choose the correct model.

### 4.4.3 High Dimensional Logistic Regression

Let's consider the previous logistic regression again,

$$\text{logit } P(y = 1|\mathbf{x}) = 2x_1 + x_2 + 3x_3$$

But this time we increase $p$ from 50 to 2000 while keeping $n$ still as 500. Now it becomes a high-dimensional problem.

For $\pi(s)$, we tried four different $\gamma$ values : $(0.5, 0.6, 0.7, 0.8)$ and fixed the upper bound $q$ as 50. For $\pi(\boldsymbol{\beta}_s)$, we let $C_0 = 10$. We implemented similar MCMC algorithm to generate posterior samples of $(s, \boldsymbol{\beta}_s)$. In total $10^6$ samples were generated, the first $10^5$ were deleted as "burn-in" and for the left, every 10th samples were finally saved. After collecting useful samples of $(s, \boldsymbol{\beta}_s)$, we calculated both AEBIC and EBIC for each model having occurred more than 20 times.

We repeated this simulation 100 times and summarized the results in the following table.

Table 4.3: Simulation results of EBIC and AEBIC for the high dimensional logistic regression

| $\gamma$ | 0.5 | 0.6 | 0.7 | 0.8 |
|---|---|---|---|---|
| Mean of Size (SD of size) | 3.60(0.96) | 3.21(0.46) | 3.15(0.43) | 3.07(0.30) |
| Positive Selection Rate | 1.000 | 1.000 | 1.000 | 1.000 |
| True Discovery Rate | 0.833 | 0.935 | 0.949 | 0.977 |
| Mean of Size (SD of size) | 3.39(0.89) | 3.11(0.31) | 3.06(0.28) | 3.02(0.20) |
| Positive Selection Rate | 1.000 | 1.000 | 1.000 | 0.997 |
| True Discovery Rate | 0.885 | 0.965 | 0.977 | 0.990 |

From table 4.3, we can see as $\gamma$ increases from 0.5 to 0.8, both EBIC and AEBIC tend to select more sparse models, which is illustrated by their decreased model sizes

and increased true discovery rates. This is due to the fact that larger $\gamma$ imposes a heavier penalty on the model size. In addition, both of their positive selection rate stay at 1 and don't decrease until $\gamma$ reaches 0.8. When $\gamma$ equals 0.8, PSR for AEBIC slightly decreases to 0.997.

### 4.4.4 High Dimensional Linear Regression

Let's consider the previous linear regression again,

$$y = x_1 + 2.2x_2 - 1.6x_3 + 2x_4 - 1.4x_5 + \epsilon$$

This time we increase $p$ from 50 to 2000 while keeping $n$ still as 500 to let it be a high-dimensional problem

For $\pi(s)$, we tried four different $\gamma$ values : $(0.5, 0.6, 0.7, 0.8)$ and fixed the upper bound $q$ as 50. For $\pi(\boldsymbol{\beta}_s)$, we let $C_0 = 10$. This time, we implemented a slightly modified MCMC algorithm to generate posterior samples of $(s, \boldsymbol{\beta}_s)$. That is, in the birth step, we no longer randomly select new variables with same probability. Instead, we assign a weight to each variable and then randomly select variables based on their weights. The weight for any $x_k$ is proportional to $|\rho(y, x_k)|$, where $\rho$ is the pearson correlation.This modification can significantly accelerate the convergence of the Markov chain. In total $10^6$ samples were generated, the first $10^5$ were deleted as "burn-in" and for the left, every 10th samples were finally saved. After collecting useful samples of $(s, \boldsymbol{\beta}_s)$, we calculated both AEBIC and EBIC for each model having occurred more than 20 times.

We repeated this simulation 100 times and summarized the results in the following table.

From table 4.4, we can observe the familiar trend. As $\gamma$ increases, both EBIC and AEBIC prefer more sparse models.

Table 4.4: Simulation results of EBIC and AEBIC for the high dimensional linear regression

| $\gamma$ | 0.5 | 0.6 | 0.7 | 0.8 |
|---|---|---|---|---|
| Mean of Size (SD of size) | 5.30(0.54) | 5.15(0.39) | 5.07(0.26) | 5.03(0.17) |
| Positive Selection Rate | 1.000 | 1.000 | 1.000 | 1.000 |
| True Discovery Rate | 0.943 | 0.971 | 0.986 | 0.994 |
| Mean of Size (SD of size) | 5.16(0.42) | 5.06(0.24) | 5.05(0.22) | 5.01(0.10) |
| Positive Selection Rate | 1.000 | 1.000 | 1.000 | 1.000 |
| True Discovery Rate | 0.969 | 0.988 | 0.990 | 0.998 |

## 4.5 Real Data Example

In Singh *et al.* (2002), the researchers measured 6033 genes on 102 samples (52 prostate cancer patients and 50 controls) with the aim of exploring the relationship between these 6033 genes and the prostate cancer. In the next several years, this dataset has been analyzed in multiple articles, such as (Chen and Chen, 2012; Efron, 2009; Liang *et al.*, 2013).

In this section, we re-analyze the dataset by using our algorithm. First, we build a logistic regression model with intercept term

$$\text{logit } P(y = 1|\mathbf{x}) = x^T\boldsymbol{\beta}$$

where $y$ denotes the status of prostate cancer, $\mathbf{x}$ includes the microarray measurements of 6033 genes and 1 for the intercept, so $n = 102$ and $p = 6034$. To identify the active features and select the most reliable model, we generate posterior samples of $(s, \boldsymbol{\beta}_s)$ by implementing the MCMC algorithms. For $\pi(s)$, we let $\gamma = 0.7$ and fix the upper bound $q$ as 50. For $\pi(\boldsymbol{\beta}_s)$, we let $C_0 = 10$. The MH move includes three types as before, that is, birth, death and parameter update. In the birth step, the

weight for gene $k$ is defined as

$$w_k = \text{Null deviance} - \text{Deviance}_k + 0.1$$

where Null deviance denotes the deviance of the model including only intercept term and $\text{Deviance}_k$ denotes the deviance of the model only including intercept term and gene $k$. In total $10^7$ samples were generated, the first $10^6$ were discarded as "burn-in" and for the left, every 10th samples were finally saved. After collecting useful samples of $(s, \boldsymbol{\beta}_s)$, we calculated AEBIC for each model having occurred more than 10 times.

In contrast to the previous simulation studies, where the second smallest AEBIC is far behind the first one, this time we observed that multiple models have very similar AEBICs. This phenomenon is reasonable. After all, this dataset only contains 102 observations, which is not large enough to guarantee the consistency property of our algorithm. Therefore, in this case, it is unappropriate to only consider the model

Table 4.5: AEBIC and 1-CVMR for the top 10 models with highest AEBIC

| Rank | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $\text{AEBIC}_{0.8}$ | 131.36 | 131.73 | 134.77 | 137.13 | 137.70 |
| 1-CVMR | 0.118 | 0.127 | 0.167 | 0.137 | 0.137 |
| Rank | 6 | 7 | 8 | 9 | 10 |
| $\text{AEBIC}_{0.7}$ | 138.11 | 138.59 | 139.19 | 139.31 | 139.65 |
| 1-CVMR | 0.078 | 0.098 | 0.186 | 0.157 | 0.107 |

with the minimum AEBIC. Instead, we should consider all models with relatively very small AEBICS, such as the top 10 models. We further calculated the "leave-one-out cross-validation misclassification rate", abbreviated by "1-CVMR", for each

Table 4.6: Top 10 genes in Chen and Chen (2012) and their corresponding rankings by our method

| Gene | 610 | 1720 | 332 | 364 | 1068 |
|---|---|---|---|---|---|
| Chen and Chen (2012)'s ranking | 1 | 2 | 3 | 4 | 5 |
| Our ranking | 1 | 15 | 8 | 6 | 10 |
| Gene | 914 | 3940 | 1077 | 4331 | 579 |
| Chen and Chen (2012)'s ranking | 6 | 7 | 8 | 9 | 10 |
| Our ranking | 5 | 3 | 7 | 4 | 19 |

of the top 10 models and listed them in table 4.5.

In addition, we also ranked genes based on their appearing frequencies in the top 100 models and it is interesting that our ranking has a lot of overlapping with the ranking in (Chen and Chen, 2012),which was established by a totally different procedure. We listed the genes ranked as top 10 in Chen and Chen (2012) and their corresponding rankings by our method in table 4.6 for reference.

# 5. SUMMARY AND DISCUSSIONS

This dissertation covers three different topics on Big data and High dimensional data.

In Chapter 2, we introduced HZ-SIS as a new model-free feature screening method, and established its sure screening property under the ultrahigh dimensional setting. The HZ-SIS method contains two components, nonparanormal transformation and HZ-test. The numerical examples indicate that, compared to the existing methods, HZ-SIS can achieve better performance when the covariates follow a heavy-tailed distribution and when the underlying true model is complex with interaction variables. The reason why HZ-SIS can achieve such a robust performance can be understood from two perspectives. First, HZ-SIS does not require any extra conditions except for two regularity conditions which are generally required for high-dimensional feature screening. Second, the truncated empirical CDF estimator used in the estimated nonparanormal transformation helps to reduce the effect of extreme data.

In HZ-SIS, the HZ-test is employed to test the normality of the nonparanormally transformed data. Other than the HZ-test, other multivariate normality tests, such as Szekeley-Rizzo's goodness-of-fit test (Székely and Rizzo, 2005) and Mardia's skewness and kurtosis tests (Mardia, 1970), can also be applied here. Since none of the tests are universally superior, a combination of different tests might produce a higher power. How to combine different tests to get a higher power test will be one of our future research topics.

Henze and Zirkler (1990) showed that under the null hypothesis that the testing data are drawn from a multivariate Gaussian distribution, the HZ-test statistic follows a log-normal distribution. This implies that $\widetilde{\omega}_k^*$ approximately follows a log-

normal distribution, although a rigorous theoretical justification is still needed to account for the effect caused by the estimation error of the nonparanormal transformation. Then, compared to the existing variable screening methods, HZ-SIS will have an added advantage that the relevance of an individual predictor to the response variable can be measured with $p$-value, and thus many of the existing multiple hypothesis tests can be applied to the problem to assist variable screening.

In Chapter 3, we developed a simple, practical and efficient MCMC algorithm for Bayesian analysis of big data. The proposed algorithm has two innovations. First, it provides a simple and practical way to aggregate subposteriors to approximate the full data posterior. Second, it suggests to implement the Pop-SAMC algorithm to simulate from each subposterior. Since the whole algorithm consists of two levels of parallel, data parallel and simulation parallel, it is called a Double Parallel Monte Carlo algorithm. Theoretically, we have shown that the double parallel algorithm can produce a good approximation to the full data posterior distribution. Empirically, we have demonstrated that the results produced by the double parallel algorithm agree well with those generated from the full data posterior, while enabling massive speed-ups in computational time.

The double parallel algorithm works based on Laplace's method, but it can also cover some problems that are traditionally treated as discrete, such as variable selection problems. As shown in Section 3.5, these problems can be treated as continuous by imposing a local shrinkage prior on the space of variable coefficients. A further extension of the proposed algorithm to general discrete parameter space will be of great interest.

In Chapter 4, we proposed ABIC, an innovative way of using posterior samples to conduct variable selection. We also established the consistency property of this information criterion for the high-dimensional generalized linear model under some

sparsity and regularity conditions.

In order to simplify the technical details, we imposed relatively stronger conditions on the true model. In fact, the consistency should still hold under weaker conditions. For example, we can extend the canonical link to non-canonical link; we can also consider models with dispersion parameter (such as linear regression with unknown variance); the upper bound $q$ of the true model size $|s^*|$ can be allowed to increase with $n$, rather than being a constant; the total number of variables $p_n$ can be also allowed to grow exponentially with $n$, instead of polynomially, etc. How to derive the theoretical proof of consistency under these weaker conditions will be one of our research goals in the future.

As mentioned before, though this information criterion should be applicable to a broader class of models, including some nonlinear models, this dissertation only considers generalized linear model. How to modify the prior settings, implementation procedures and consistency analysis to deal with nonlinear models would be a very challenging topic and also one of our research goals in the future.

Another future work may be related to small sample problem. We noticed that both the derivation and theoretical properties of this method are based on the large sample assumption. When the sample size is relatively small, such as the dataset used in section 5, the performance of this method is less satisfactory. In fact, this problem also exists in most of high-dimensional variable selection methods. Therefore, how to improve the performance of model selection for high-dimensional regression with small sample size, would be of great interest.

# REFERENCES


Abramovich, F., Benjamini, Y., Donoho, D. L., and Johnstone, I. M. (**2006**). "Adapting to unknown sparsity by controlling the false discovery rate", Ann. Statist. **34**, 584–653.

Akaike, H. (**1974**). "A new look at the statistical model identification", Automatic Control, IEEE Transactions **19**, 716–723.

Andrieu, C., Moulines, E., and Priouret, P. (**2005**). "Stability of stochastic approximation under verifiable conditions", SIAM Journal of Control and Optimization **44**, 283–312.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A., Kim, S., Wilson, C., Lehar, J., Kryukov, G., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M., Monahan, J., Morais, P., Meltzer, J., Korejwa, A., Jane-Valbuena, J. andMapa, F., Thibault, J., Bric-Furlong, E., Raman, P., Shipway, A., Engels, I., and et al. (**2012**). "The Cancer Cell Line Encyclopedia enables predictive modeling of anticancer drug sensitivity", Nature **483**, 603–607.

Buzdar, A. (**2009**). "Role of biologic therapy and chemotherapy in hormone receptor and HER2-positive breast cancer", The Annals of Oncology **20**, 993–999.

Candes, E. and Tao, T. (**2007**). "The Dantzig selector: statistical estimation when p is much larger than n", Ann. Statist. **35**, 2313–2351.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (**2010**). "The horseshoe estimator for sparse signals", Biometrika **97**, 465–480.

Chen, J. and Chen, Z. (**2008**). "Extended Bayesian information criteria for model selection with large model spaces", Biometrika **95**, 759–771.

Chen, J. and Chen, Z. (**2012**). "Extended BIC for small-n-large-p sparse GLM",

Statistica Sinica **22**, 555–574.

Chen, Z. and Luo, S. (**2013**). "Selection consistency of EBIC for GLIM with non-canonical links and diverging number of parameters", Statistics and Its Interface **6**, 275–284.

Cui, H., Li, R., and Zhong, W. (**2015**). "Model-free feature screening for ultrahigh dimensional discriminant analysis", Journal of the American Statistical Association **110**, 630–641.

Efron, B. (**2009**). "Empirical Bayes estimates for large-scale prediction problems", Journal of the American Statistical Association **104**, 1015–1028.

Fan, J., Feng, Y., and Song, R. (**2011**). "Nonparametric independence screening in sparse ultra-high-dimensional additive models", Journal of the American Statistical Association **106**, 544–557.

Fan, J., Han, F., and Liu, H. (**2014**). "Challenges of big data analysis", National Science Review **1**, 293–314.

Fan, J. and Li, R. (**2001**). "Variable selection via nonconcave penalized likelihood and its oracle properties", Journal of the American Statistical Association **96**, 1348–1360.

Fan, J. and Li, R. (**2007**). "Statistical challenges with high dimensionality: feature selection in knowledge discovery", in *Proceedings of the International Congress of Mathematicians Madrid, August 22-30, 2006*, 595–622.

Fan, J. and Lv, J. (**2008**). "Sure independence screening for ultrahigh dimensional feature space", Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70**, 849–911.

Fan, J. and Lv, J. (**2010**). "A selective overview of variable selection in high dimensional feature space", Statistica Sinica **20**, 101–148.

Fan, J., Samworth, R., and Wu, Y. (**2009**). "Ultrahigh dimensional feature selection:

Beyond the linear model", J. Mach. Learn. Res. **10**, 2013–2038.

Fan, J. and Song, R. (**2010**). "Sure independence screening in generalized linear models with NP-dimensionality", Ann. Statist. **38**, 3567–3604.

Foster, D. P. and George, E. I. (**1994**). "The risk inflation criterion for multiple regression", Ann. Statist. **22**, 1947–1975.

Foygel, R. and Drton, M. (**2011**). "Bayesian model choice and information criteria in sparse generalized linear models", ArXiv 1112.5635 .

Foygel Barber, R., Drton, M., and Tan, K. M. (**2015**). "Laplace approximation in high-dimensional Bayesian regression", ArXiv 1503.08337 .

Geman, S. and Geman, D. (**1984**). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images", IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721–741.

Grünwald, V. and Hidalgo, M. (**2003**). "Developing inhibitors of the epidermal growth factor receptor for cancer treatment", Journal of the National Cancer Institute **95**, 851–867.

Hadley, K. E. and Hendricks, D. T. (**2014**). "Use of NQO1 status as a selective biomarker for oesophageal squamous cell carcinomas with greater sensitivity to 17-AAG", BMC Cancer **14**, 1–8.

Hall, P. and Miller, H. (**2009**). "Using generalized correlation to effect variable selection in very high dimensional problems", Journal of Computational and Graphical Statistics **18**, 533–550.

Hasting, W. (**1970**). "Monte Carlo sampling methods using Markov chains and their applications", Biometrika **57**, 97–109.

Haughton, D. M. A. (**1988**). "On the choice of a model to fit data from an exponential tmily", Ann. Statist. **16**, 342–355.

He, X., Wang, L., and Hong, H. G. (**2013**). "Quantile-adaptive model-free variable

screening for high-dimensional heterogeneous data", Ann. Statist. **41**, 2699.

Henze, N. and Zirkler, B. (**1990**). "A class of invariant consistent tests for multivariate normality", Communications in statistics - Theory and Methods **10**, 3595–3617.

Huang, J., Horowitz, J. L., and Ma, S. (**2008**). "Asymptotic properties of bridge estimators in sparse high-dimensional regression models", Ann. Statist. **36**, 587–613.

Jiang, B. and Liu, J. (**2014**). "Sliced inverse regression with variable selection and interaction detection ", ArXiv 1304.4056 .

Johnson, V. E. and Rossell, D. (**2012**). "Bayesian model selection in high-dimensional settings", Journal of the American Statistical Association **107**, 649–660.

Johnstone, I. M. and Titterington, D. M. (**2009**). "Statistical challenges of high-dimensional data", Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. **367**, 4237–4253.

Kass, R. E., Tierney, L., and Kadane, J. B. (**1990**). "The validity of posterior expansions based on Laplace's method", in *Bayesian and likelihood methods in statistics and econometrics: essays in honor of George A. Barnard*, edited by S. Geisser, J. S. Hodges, S. J. Press, and A. ZeUner, volume 7, 473–488.

Li, R., Zhong, W., and Zhu, L. (**2012**). "Feature screening via distance correlation learning", Journal of the American Statistical Association **107**, 1129–1139.

Liang, F. (**2009**). "On the use of stochastic approximation Monte Carlo for Monte Carlo integration", Statistics & Probability Letters **79**, 581 – 587.

Liang, F., Kim, J., and Song, Q. (**2016**). "A Bootstrap Metropolis-Hastings algorithm for Bayesian analysis of big data", Technometrics **58**, 304–318.

Liang, F., Liu, C., and Carroll, R. J. (**2007**). "Stochastic approximation in Monte Carlo computation", Journal of the American Statistical Association **102**, 305–320.

Liang, F., Song, Q., and Qiu, P. (**2015**). "An equivalent measure of partial correla-

tion coefficients for high-dimensional Gaussian graphical models", Journal of the American Statistical Association **110**, 1248–1265.

Liang, F., Song, Q., and Yu, K. (**2013**). "Bayesian subset modeling for high-dimensional generalized linear models", Journal of the American Statistical Association **2013**, 589–606.

Liu, H., Lafferty, J., and Wasserman, L. (**2009**). "The nonparanormal: Semiparametric estimation of high dimensional undirected graphs", J. Mach. Learn. Res. **10**, 2295–2328.

Mallows, C. L. (**1973**). "Some comments on Cp", Technometrics **15**, 661–675.

Mardia, K. (**1970**). "Measures of multivariate skewness and kurtosis with applications", Biometrika **57**, 519–530.

McCullagh, P. and Nelder, J. A. (**1989**). *Generalized linear models (Second edition)* (London: Chapman & Hall).

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (**1953**). "Equation of state calculations by fast computing machines", The Journal of Chemical Physics **21**, 1087–1092.

Minsker, S., Srivastava, S., Lin, L., and Dunson, D. B. (**2014**). "Robust and scalable Bayes via a median of subset posterior measures", ArXiv 1403.2660 .

Neiswanger, W., Wang, C., and Xing, E. (**2013**). "Asymptotically exact, embarrassingly parallel MCMC", ArXiv 1311.4780 .

Nishii, R. (**1984**). "Asymptotic properties of criteria for selection of variables in multiple regression", Ann. Statist. **12**, 758–765.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (**1997**). "Bayesian model averaging for linear regression models", Journal of the American Statistical Association **92**, 179–191.

Ripley, B. D. (**1987**). *Stochastic Simulation* (John Wiley & Sons, Inc., New York,

NY, USA).

Robbins, H. and Monro, S. (**1951**). "A stochastic approximation method", Ann. Math. Statist. **22**, 400–407.

Roe, B. P., Yang, H.-J., Zhu, J., Liu, Y., Stancu, I., and McGregor, G. (**2005**). "Boosted decision trees as an alternative to artificial neural networks for particle identification", Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **543**, 577 – 584.

Schwarz, G. (**1978**). "Estimating the dimension of a model", Ann. Statist. **6**, 461–464.

Scott, S. L., Blocker, A. W., and Bonassi, F. V. (**2016**). "Bayes and big data: The consensus Monte Carlo algorithm", International Journal of Management Science and Engineering Management **11**, 78–88.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R., and Sellers, W. R. (**2002**). "Gene expression correlates of clinical prostate cancer behavior", Cancer Cell **1**, 203 – 209.

Song, Q. and Liang, F. (**2015**). "High-dimensional variable selection with reciprocal L1-regularization", Journal of the American Statistical Association **110**, 1607–1620.

Song, Q., Wu, M., and Liang, F. (**2014**). "Weak convergence rates of population versus single-chain stochastic approximation MCMC algorithms", Adv. in Appl. Probab. **46**, 1059–1083.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (**2002**). "Bayesian measures of model complexity and fit", Journal of the Royal Statistical Society Series B **64**, 583–639.

Srivastava, S., Li, C., and Dunson, D. B. (**2015**). "Scalable Bayes via Barycenter in Wasserstein space", ArXiv 1508.05880 .

Stone, M. (**1974**). "Cross-validatory choice and assessment of statistical predictions (with discussion)", Journal of the Royal Statistical Society, Series B: Methodological **36**, 111–147.

Székely, G. and Rizzo, M. (**2005**). "A new test for multivariate normality", Journal of Multivariate Analysis **93**, 58–80.

Székely, G., Rizzo, M., and Bakirov, N. (**2007**). "Measuring and testing dependence by correlation of distances", Annals of Statistics **35**, 2769–2794.

Tibshirani, R. (**1996**). "Regression shrinkage and selection via the Lasso", Journal of the Royal Statistical Society (Series B) **58**, 267–288.

Wahba, G., C. P. (**1979**). "Smoothing noisy data with spline functions. estimating the correct degree of smoothing by the method of generalized cross-validation.", Numerische Mathematik **31**, 377–404.

Wang, K., Shrestha, R., Wyatt, A. W., Reddy, A., Wang, Y., Lapuk, A., and Collins, C. C. (**2014**). "A meta-analysis approach for characterizing pan-cancer mechanisms of drug sensitivity in cell lines", PLoS One **9**, 1–16.

Wang, X. and Dunson, D. (**2013**). "Parallelizing MCMC via Weierstrass sampler", ArXiv 1312.4605 .

Zhang, C.-H. (**2010**). "Nearly unbiased variable selection under minimax concave penalty", Ann. Statist. **38**, 894–942.

Zhu, L.-P., Li, L., Li, R., and Zhu, L.-X. (**2011**). "Model-free feature screening for ultrahigh-dimensional data", Journal of the American Statistical Association **106**, 1464–1475.

Zoppoli, G., Regairaz, M., Leo, E., Reinhold, W., Varma, S., Ballestrero, A., Doroshow, J., and Pommier, Y. (**2012**). "Putative DNA/RNA helicase schlafenl1 (SLFN11) sensitizes cancer cells to DNA-damaging agents", Proceedings of the National Academy of Sciences USA **109**, 15030–15035.

Zou, H. and Hastie, T. (**2005**). "Regularization and variable selection via the elastic net", Journal of the Royal Statistical Society. Series B (Statistical Methodology) **67**, 301–320.

APPENDIX A

SUPPLEMENTARY MATERIALS FOR CHAPTER 2

## A.1 Proof of Lemma 2.1

Define

$$\omega_k^* = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} e^{-\frac{\beta^2}{2}E_{ij}} - \frac{2}{n(1+\beta^2)}\sum_{i=1}^{n} e^{-\frac{\beta^2}{2(1+\beta^2)}E_i} + \frac{1}{1+2\beta^2},$$

where

$$\begin{aligned} E_{ij} &= (\Phi^{-1}(F_k(x_{ki})) - \Phi^{-1}(F_k(x_{kj})))^2 + (\Phi^{-1}(F_y(y_i)) - \Phi^{-1}(F_y(y_j)))^2, \\ E_i &= \Phi^{-1}(F_k(x_{ki}))^2 + \Phi^{-1}(F_y(y_i))^2. \end{aligned}$$

For any $\varepsilon > 0$, we have

$$\begin{aligned} P(|\widetilde{\omega}_k^* - \omega_k| > \varepsilon) &= P(|\widetilde{\omega}_k^* - \omega_k^* + \omega_k^* - \omega_k| > \varepsilon) \\ &\leq P(|\widetilde{\omega}_k^* - \omega_k^*| > \frac{\varepsilon}{2}) + P(|\omega_k^* - \omega_k| > \frac{\varepsilon}{2}). \end{aligned}$$

For simplicity, in what follows we let $\widetilde{T}_k(x) \equiv \Phi^{-1}(\widetilde{F}_k(x))$, $T_k(x) \equiv \Phi^{-1}(F_k(x))$, and $g_j \equiv T_j^{-1}$.

For the first term, we have

$$
\begin{aligned}
P(|\widetilde{\omega}_k^* - \omega_k^*| > \frac{\varepsilon}{2}) &\leq P(\frac{1}{n^2}\sum_{i,j}|e^{-\frac{\beta^2}{2}Dij} - e^{-\frac{\beta^2}{2}Eij}| > \frac{\varepsilon}{4}) \\
&\quad + P(\frac{1}{n}\sum_{i}|e^{-\frac{\beta^2}{2(1+\beta^2)}Di} - e^{-\frac{\beta^2}{2(1+\beta^2)}Ei}| > \frac{1+\beta^2}{2}\frac{\varepsilon}{4}) \\
&\leq P(\frac{1}{n^2}\sum_{i,j}\frac{\beta^2}{2}|D_{ij} - E_{ij}| > \frac{\varepsilon}{4}) + P(\frac{1}{n}\sum_{i}\frac{1}{2}|D_i - E_i| > \frac{1+\beta^2}{2}\frac{\varepsilon}{4}) \\
&= P(\frac{1}{n^2}\sum_{i,j}|D_{ij} - E_{ij}| > \frac{\varepsilon}{2\beta^2}) + P(\frac{1}{n}\sum_{i}|D_i - E_i| > \frac{(1+\beta^2)\varepsilon}{4}).
\end{aligned}
$$

Note that we only deal with $P(\frac{1}{n^2}\sum_{i,j}|D_{ij}-E_{ij}| > \frac{\varepsilon}{2\beta^2})$, because $P(\frac{1}{n}\sum_i|D_i-E_i| > \frac{(1+\beta^2)\varepsilon}{4})$ can be calculated in a similar way. First, we calculate

$$
\begin{aligned}
& D_{ij} - E_{ij} \\
= \; & (\widetilde{T}_k(x_{ki}) - \widetilde{T}_k(x_{kj}))^2 + (\widetilde{T}_y(y_i) - \widetilde{T}_y(y_j))^2 \\
& -(T_k(x_{ki}) - T_k(x_{kj}))^2 - (T_y(y_i) - T_y(y_j))^2 \\
= \; & (\widetilde{T}_k(x_{ki}) - T_k(x_{ki}))^2 + (\widetilde{T}_k(x_{kj}) - T_k(x_{kj}))^2 - 2(\widetilde{T}_k(x_{ki})\widetilde{T}_k(x_{kj}) - T_k(x_{ki})T_k(x_{kj})) \\
& +(\widetilde{T}_y(y_i) - T_y(y_i))^2 + (\widetilde{T}_y(y_j) - T_y(y_j))^2 - 2(\widetilde{T}_y(y_i)\widetilde{T}_y(y_j) - T_y(y_i)T_y(y_j)) \\
= \; & (\widetilde{T}_k(x_{ki}) - T_k(x_{ki}))(\widetilde{T}_k(x_{ki}) + T_k(x_{ki})) + (\widetilde{T}_k(x_{kj}) - T_k(x_{kj}))(\widetilde{T}_k(x_{kj}) + T_k(x_{kj})) \\
& -2(\widetilde{T}_k(x_{ki}) - T_k(x_{ki}))(\widetilde{T}_k(x_{kj}) - T_k(x_{kj})) - 2(\widetilde{T}_k(x_{ki}) - T_k(x_{ki}))T_k(x_{kj}) \\
& -2(\widetilde{T}_k(x_{kj}) - T_k(x_{kj\varsigma}))T_k(x_{ki}) + (\widetilde{T}_y(y_i) - T_y(y_i))(\widetilde{T}_y(y_i) + T_y(y_i)) \\
& +(\widetilde{T}_y(y_j) - T_y(y_j))(\widetilde{T}_y(y_j) + T_y(y_j)) - 2(\widetilde{T}_y(y_i) - T_y(y_i))(\widetilde{T}_y(y_j) - T_y(y_j)) \\
& -2(\widetilde{T}_y(y_i) - T_y(y_i))T_y(y_j) - 2(\widetilde{T}_y(y_j) - T_y(y_j))T_y(y_i).
\end{aligned}
$$

Among the ten terms, $(\widetilde{T}_k(x_{ki})-T_k(x_{ki}))(\widetilde{T}_k(x_{kj})-T_k(x_{kj}))$ and $(\widetilde{T}_y(y_i)-T_y(y_i))(\widetilde{T}_y(y_j)-T_y(y_j))$ are of a higher order, and the other terms share the same order. Hence, we

only consider the probability $P(\frac{1}{n^2}\sum_{i,j}|(\widetilde{T}_y(y_i)-T_y(y_i))T_y(y_j)|>\frac{\varepsilon}{20\beta^2})$.

Define the event $\mathbb{A}_n$ as

$$\mathbb{A}_n \equiv \{g_y(-n^d) \le y_1, \ldots, y_n \le g_y(n^d)\}.$$

Since for the standard Gaussian random variable $Z$,

$$P(Z>t) \le \frac{1}{\sqrt{2\pi}}\frac{e^{-\frac{1}{2}t^2}}{t}, \quad \text{if } t>1, \tag{A.1}$$

we have

$$P(\mathbb{A}_n^c) \le \sum_{i=1}^{n} 2P(y_i > g_y(n^d)) \le 2n\frac{1}{\sqrt{2\pi}}\frac{e^{-\frac{1}{2}n^{2d}}}{n^d} = \sqrt{\frac{2}{\pi}}n^{1-d}\exp\{-\frac{1}{2}n^{2d}\}.$$

Therefore,

$$\begin{aligned} P(\frac{1}{n^2} \quad \sum_{i,j} \quad & |(\widetilde{T}_y(y_i)-T_y(y_i))T_y(y_j)|>\frac{\varepsilon}{20\beta^2}) \\ \le \quad & P(\frac{1}{n^2}\sum_{i,j}|(\widetilde{T}_y(y_i)-T_y(y_i))T_y(y_j)|>\frac{\varepsilon}{20\beta^2}, \mathbb{A}_n) + P(\mathbb{A}_n^c) \\ \le \quad & P(\frac{1}{n^2}\sum_{i,j}|(\widetilde{T}_y(y_i)-T_y(y_i))T_y(y_j)|>\frac{\varepsilon}{20\beta^2}, \mathbb{A}_n) + \sqrt{\frac{2}{\pi}}n^{1-d}\exp\{-\frac{1}{2}n^{2d}\}. \end{aligned}$$

For simplicity, henceforth, we let $\Delta_{ij} = (\widetilde{T}_y(y_i)-T_y(y_i))T_y(y_j)$.

Set the truncation parameter $\delta_n = \frac{1}{4n^{\frac{m}{2}}\sqrt{2\pi m\log n}}$, $m<1$ and split the interval $[g_y(-n^d), g_h(n^d)]$ into

$$\mathbb{M}_n = (g_y(-\sqrt{m\log n}), g_h(\sqrt{m\log n}))$$

and

$$\mathbb{E}_n = [g_y(-n^d), g_y(-\sqrt{m \log n})] \cup [(g_h(\sqrt{m \log n}), g_y(n^d)].$$

Therefore,

$$\begin{aligned} P(\frac{1}{n^2}\sum_{i,j}|\Delta_{ij}| > \frac{\varepsilon}{20\beta^2}, \mathbb{A}_n) &\leq P(\frac{1}{n^2}\sum_{y_i \in \mathbb{E}_n \cup y_j \in \mathbb{E}_n}|\Delta_{ij}| > \frac{\varepsilon}{40\beta^2}) \\ &\quad + P(\frac{1}{n^2}\sum_{y_i \in \mathbb{M}_n \cap y_j \in \mathbb{M}_n}|\Delta_{ij}| > \frac{\varepsilon}{40\beta^2}). \end{aligned}$$

We now analyze these two terms separately.

From Lemma 12.3 of Abramovich *et al.* (2006), if we let $\Phi^{-1}(\eta)$ denote the upper $\eta$th percentile of the standard Gaussian distribution, for $\eta \geq 0.99$ we have

$$\Phi^{-1}(\eta) = \sqrt{2 \log \frac{1}{1-\eta}} - r(\eta), \qquad r(\eta) \in [0, 1.5].$$

Based on this lemma, we can show

$$\begin{aligned} |\widetilde{T}_y(t)| < \Phi^{-1}(1-\delta_n) &= \sqrt{2 \log \frac{1}{\delta_n}} - r(1-\delta_n) \\ &\leq \sqrt{2[\frac{m}{2}\log(n) + \log(4\sqrt{2\pi m \log n})]} < \sqrt{\log n}, \end{aligned}$$

for any $t \in \mathbb{R}$, provided that $n$ is sufficiently large.

Then we can bound $\Delta_{ij}$ under $\mathbb{A}_n$:

$$\begin{aligned} |\Delta_{ij}| = |\widetilde{T}_y(y_i) - T_y(y_i)||T_y(y_j)| &\leq (|\widetilde{T}_y(y_i)| + |T_y(y_i)|)|T_y(y_j)| \\ &\leq (\sqrt{\log(n)} + n^d)n^d < 2n^{2d}, \end{aligned}$$

if $n$ is sufficiently large. Therefore,

$$\begin{aligned}
& P(\frac{1}{n^2}\sum_{y_i\in\mathbb{E}_n\cup y_j\in\mathbb{E}_n}|\Delta_{ij}|>\frac{\varepsilon}{40\beta^2}) \\
\leq\ & P(\frac{1}{n^2}\sum_{i,j}\mathbf{1}_{\{y_i\in\mathbb{E}_n\cup y_j\in\mathbb{E}_n\}}>\frac{\varepsilon}{80n^{2d}\beta^2}) \\
\leq\ & P(\frac{1}{n^2}\sum_{i,j}\mathbf{1}_{\{y_i\in\mathbb{E}_n\}}+\mathbf{1}_{\{y_j\in\mathbb{E}_n\}}>\frac{\varepsilon}{80n^{2d}\beta^2}) \\
=\ & P(\frac{1}{n}\sum_{i}\mathbf{1}_{\{y_i\in\mathbb{E}_n\}}>\frac{\varepsilon}{160n^{2d}\beta^2}) \\
\leq\ & P(\frac{1}{n}\sum_{i}(\mathbf{1}_{\{y_i\in\mathbb{E}_n\}}-P(y_i\in\mathbb{E}_n))>\frac{\varepsilon}{160n^{2d}\beta^2}-n^{-\frac{m}{2}}),
\end{aligned}$$

where the last inequality follows from the fact that

$$P(y_i\in\mathbb{E}_n)\leq 2P(y_i>g_h(\sqrt{m\log n}))\leq 2\frac{1}{\sqrt{2\pi}}\frac{e^{-\frac{1}{2}m\log n}}{\sqrt{m\log n}}\leq n^{-\frac{m}{2}}.$$

Recall $\varepsilon=cn^{-\kappa}$ and assume $\kappa+2d<\frac{m}{2}$, we have

$$\frac{\varepsilon}{160n^{2d}\beta^2}-n^{-\frac{m}{2}}=\frac{cn^{-(\kappa+2d)}}{160\beta^2}-n^{-\frac{m}{2}}\geq\frac{cn^{-(\kappa+2d)}}{200\beta^2},$$

if $n$ is sufficiently large. Further, we have

$$\begin{aligned}
& P(\frac{1}{n}\sum_{i}(\mathbf{1}_{\{y_i\in\mathbb{E}_n\}}-P(y_i\in\mathbb{E}_n))>\frac{\varepsilon}{160n^{2d}\beta^2}-n^{-\frac{m}{2}}) \\
& \leq P(\frac{1}{n}\sum_{i}(\mathbf{1}_{\{y_i\in\mathbb{E}_n\}}-P(y_i\in\mathbb{E}_n))>\frac{cn^{-(\kappa+2d)}}{200\beta^2}) \\
& \leq \exp\{-2n\frac{c^2n^{-2(\kappa+2d)}}{40000\beta^4}\}=\exp\{-\frac{c^2}{20000\beta^4}n^{1-2(\kappa+2d)}\},
\end{aligned}$$

where the last inequality follows from Hoeffiding's inequality.

Now we turn to $P(\frac{1}{n^2}\sum_{y_i\in\mathbb{M}_n\cap y_j\in\mathbb{M}_n}|\Delta_{ij}|>\frac{\varepsilon}{40\beta^2})$. Define the event $\mathbb{B}_n$ as

$$\mathbb{B}_n \equiv \{\delta_n \le \widehat{F}_y(g_j(-\sqrt{\beta\log n}))\cup \widehat{F}_y(g_j(\sqrt{\beta\log n}))\le 1-\delta_n\}.$$

Following from (A.1), we have

$$\begin{aligned}
& P(\mathbb{B}_n^c)\\
\le\ & 2P(\widehat{F}_y(g_j(\sqrt{\beta\log n}))\ge 1-\delta_n)\\
=\ & 2P(\widehat{F}_y(g_j(\sqrt{\beta\log n})-F_y(\sqrt{\beta\log n})\ge 1-F_y(\sqrt{\beta\log n})-\delta_n)\\
\le\ & 2P(\widehat{F}_y(g_j(\sqrt{\beta\log n})-F_y(\sqrt{\beta\log n})\ge \frac{1}{2n^{\frac{m}{2}}\sqrt{2\pi m\log n}}-\delta_n)\\
\le\ & 2P(\widehat{F}_y(g_j(\sqrt{\beta\log n})-F_y(\sqrt{\beta\log n})\ge \frac{1}{2n^{\frac{m}{2}}\sqrt{2\pi m\log n}}-\frac{1}{4n^{\frac{m}{2}}\sqrt{2\pi m\log n}})\\
=\ & 2P(\widehat{F}_y(g_j(\sqrt{\beta\log n})-F_y(\sqrt{\beta\log n})\ge \frac{1}{4n^{\frac{m}{2}}\sqrt{2\pi m\log n}})\\
\le\ & 2\exp\{-2n\frac{1}{16n^m 2\pi m\log n}\}=2\exp\{-\frac{n^{1-m}}{16\pi m\log n}\},
\end{aligned}$$

where last inequality follows from Hoeffiding's inequality. Therefore

$$\begin{aligned}
& P(\frac{1}{n^2}\sum_{y_i\in\mathbb{M}_n\cap y_j\in\mathbb{M}_n}|\Delta_{ij}|>\frac{\varepsilon}{40\beta^2})\\
&= P(\frac{1}{n^2}\sum_{y_i\in\mathbb{M}_n\cap y_j\in\mathbb{M}_n}|\Delta_{ij}|>\frac{\varepsilon}{40\beta^2},\mathbb{B}_n)+P(\frac{1}{n^2}\sum_{y_i\in\mathbb{M}_n\cap y_j\in\mathbb{M}_n}|\Delta_{ij}|>\frac{\varepsilon}{40\beta^2},\mathbb{B}_n^c)\\
&\le P(\frac{1}{n^2}\sum_{y_i\in\mathbb{M}_n\cap y_j\in\mathbb{M}_n}|\Delta_{ij}|>\frac{\varepsilon}{40\beta^2},\mathbb{B}_n)+P(\mathbb{B}_n^c)\\
&\le P(\frac{1}{n^2}\sum_{y_i\in\mathbb{M}_n\cap y_j\in\mathbb{M}_n}|\Delta_{ij}|>\frac{\varepsilon}{40\beta^2},\mathbb{B}_n)+2\exp\{-\frac{n^{1-m}}{16\pi m\log n}\}\\
&\le P(\frac{1}{n^2}\sum_{y_i\in\mathbb{M}_n\cap y_j\in\mathbb{M}_n}|\widetilde{T}_y(y_i)-T_y(y_i)|>\frac{\varepsilon}{40\beta^2\sqrt{m\log n}},\mathbb{B}_n)+2\exp\{-\frac{n^{1-m}}{16\pi m\log n}\}\\
&\le P(\sup_{t\in\mathbb{M}_n}|\widetilde{T}_y(t)-T_y(t)|>\frac{\varepsilon}{40\beta^2\sqrt{m\log n}},\mathbb{B}_n)+2\exp\{-\frac{n^{1-m}}{16\pi m\log n}\}.
\end{aligned}$$

Recall that under $\mathbb{B}_n$, for $t \in \mathbb{M}_n$, we have $\widetilde{F}_y(t) = \widehat{F}_y(t)$. So we can rewrite $\widetilde{T}_y(t) - T_y(t)$ as $\Phi^{-1}(\widehat{F}_y(t)) - \Phi^{-1}(F_y(t))$. By the mean value theorem, we further have

$$\Phi^{-1}(\widehat{F}_y(t)) - \Phi^{-1}(F_y(t)) = (\Phi^{-1})'(s)(\widehat{F}_y(t) - F_y(t)),$$

where $s$ is between $\widehat{F}_y(t)$ and $F_y(t)$. From Lemma 12.3 of Abramovich *et al.* (2006), we know $(\Phi^{-1})'(s) = \frac{1}{\phi(\Phi^{-1}(s))}$. Also, recall that under $\mathbb{B}_n$, for $t \in \mathbb{M}_n$, both $\widehat{F}_y(t)$ and $F_y(t)$ are bounded by $[\delta_n, 1-\delta_n]$. Therefore,

$$\sup_{s\in[\delta_n,1-\delta_n]} (\Phi^{-1})'(s) = \sup_{s\in[\delta_n,1-\delta_n]} \frac{1}{\phi(\Phi^{-1}(s))} = \frac{1}{\phi(\Phi^{-1}(1-\delta_n))} \leq \frac{1}{\phi(\sqrt{2\log(\frac{1}{\delta_n})})} = \frac{\sqrt{2\pi}}{\delta_n}.$$

Combining them together, we are able to show

$$\begin{aligned}
&P(\sup_{t\in\mathbb{M}_n} |\widetilde{T}_y(t) - T_y(t)| > \frac{\varepsilon}{40\beta^2\sqrt{m\log n}}, \mathbb{B}_n) \\
&\leq P(\sup_{t\in\mathbb{M}_n} |\Phi^{-1}(\widehat{F}_y(t)) - \Phi^{-1}(F_y(t))| > \frac{\varepsilon}{40\beta^2\sqrt{m\log n}}, \mathbb{B}_n) \\
&\leq P(\sup_{s\in[\delta_n,1-\delta_n]} \frac{1}{\phi(\Phi^{-1}(s))} \sup_{t\in\mathbb{M}_n} |\widehat{F}_y(t) - F_y(t)| > \frac{\varepsilon}{40\beta^2\sqrt{m\log n}}, \mathbb{B}_n) \\
&= P(\frac{\sqrt{2\pi}}{\delta_n} \sup_{t\in\mathbb{M}_n} |\widehat{F}_y(t) - F_y(t)| > \frac{\varepsilon}{40\beta^2\sqrt{m\log n}}, \mathbb{B}_n) \\
&= P(\sup_{t\in\mathbb{M}_n} |\widehat{F}_y(t) - F_y(t)| > \frac{\varepsilon}{40\beta^2\sqrt{m\log n}} \frac{1}{\sqrt{2\pi}4n^{\frac{m}{2}}\sqrt{2\pi m\log n}}, \mathbb{B}_n) \\
&= P(\sup_{t\in\mathbb{M}_n} |\widehat{F}_y(t) - F_y(t)| > \frac{\varepsilon n^{-\frac{m}{2}}}{320\beta^2\pi m\log n}, \mathbb{B}_n).
\end{aligned}$$

Using the Dvoretzky-Kiefer-Wolfowitz inequality, we have

$$\begin{aligned}
P(\sup_{t\in\mathbb{M}_n} |\widehat{F}_y(t) - F_y(t)| > \frac{\varepsilon n^{-\frac{m}{2}}}{320\beta^2\pi m\log n}, \mathbb{B}_n) &\leq \exp\{-2n[\frac{\varepsilon n^{-\frac{m}{2}}}{320\beta^2\pi m\log n}]^2\} \\
&= \exp\{-2\frac{\varepsilon^2 n^{1-m}}{102400\beta^4 m^2\pi^2\log^2 n}\}.
\end{aligned}$$

Now it remains to deal with $P(|\omega_k^* - \omega_k| > \frac{\varepsilon}{2})$. Recall that $\omega_k^*$ is a V-statistic, and the corresponding U-statistic $\omega_k^{**}$ is given by

$$\omega_k^{**} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} e^{-\frac{\beta^2}{2} E_{ij}} - \frac{2}{n(1+\beta^2)} \sum_{i=1}^{n} e^{-\frac{\beta^2}{2(1+\beta^2)} E_i} + \frac{1}{1+2\beta^2}.$$

By noting that $e^{-\frac{\beta^2}{2} E_{ij}} = 1$ when $i = j$, and $e^{-\frac{\beta^2}{2} E_{ij}} < 1$ for other cases, it is easy to show

$$\begin{aligned} |\omega_k^* - \omega_k^{**}| &= |\frac{1}{n^2}(n + \sum_{i=1}^{n} \sum_{j \neq i} e^{-\frac{\beta^2}{2} E_{ij}}) - \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} e^{-\frac{\beta^2}{2} E_{ij}}| \\ &= |\frac{1}{n} - \frac{1}{n^2(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} e^{-\frac{\beta^2}{2} E_{ij}}| \leq \frac{1}{n} + \frac{1}{n} = \frac{2}{n}. \end{aligned}$$

Recall that $\epsilon$ is of a lower order of $\frac{1}{n}$, therefore, we can consider $P(|\omega_k^{**} - \omega_k| > \frac{\varepsilon}{2})$ instead.

The kernel $h(x_{ki}, x_{kj}, y_i, y_j)$ of $\omega_k^{**}$ is bounded,

$$\begin{aligned} |h(x_{ki}, x_{kj}, y_i, y_j)| &= |e^{-\frac{\beta^2}{2} E_{ij}} - \frac{1}{1+\beta^2} e^{-\frac{\beta^2}{2(1+\beta^2)} E_i} - \frac{1}{1+\beta^2} e^{-\frac{\beta^2}{2(1+\beta^2)} E_j} + \frac{1}{1+2\beta^2}| \\ &\leq 1 + \frac{2}{1+\beta^2} + \frac{1}{1+2\beta^2} \leq 4. \end{aligned}$$

Therefore, we have

$$P(|\omega_k^{**} - \omega_k| > \frac{\varepsilon}{2}) \leq \exp\{-2[\frac{n}{2}]\frac{\varepsilon^2}{2^2}\frac{1}{8^2}\} = \exp\{-[\frac{n}{2}]\frac{\varepsilon^2}{128}\},$$

where $[\frac{n}{2}]$ denotes the integer part of $\frac{n}{2}$.

In summary, by letting $\varepsilon = cn^{-\kappa}$, we have

$$P(|\widetilde{\omega}_k^* - \omega_k| > cn^{-\kappa}) \leq O\{\exp\{-c_1 n^{2d}\} + \exp\{-c_2 n^{1-2(\kappa+2d)}\} + \exp\{-c_3 n^{1-m-2\kappa}\}\},$$

with the additional constraint $\kappa + 2d < \frac{m}{2}$. To optimize the convergence rate, we should let $m = 2(\kappa + 2d)$ and $d = \frac{1}{6} - \frac{2}{3}\kappa$, then we can obtain

$$P(|\widetilde{\omega}_k^* - \omega_k| > cn^{-\kappa}) \leq O\{\exp\{-c_1 n^{\frac{1-4\kappa}{3}}\}\}.$$

Hence,

$$P(\max_{1\leq k\leq p} |\widetilde{\omega}_k^* - \omega_k| > cn^{-\kappa}) \leq p \max_{1\leq k\leq p} P(|\widetilde{\omega}_k^* - \omega_k| > cn^{-\kappa}) \leq O\{p\exp\{-c_1 n^{\frac{1-4\kappa}{3}}\}\},$$

which completes the proof.

APPENDIX B

SUPPLEMENTARY MATERIALS FOR CHAPTER 3

## B.1 Proof of Theorem 3.1

Under conditions $(A1)$ and $(A2)$, we can expand the mean of each subposterior at the corresponding MLE, $\hat{\boldsymbol{\theta}}^{(j)}$, as follows:

$$\boldsymbol{\mu}^{(j)} = E_{\tilde{\pi}_j}(\boldsymbol{\theta}) = \hat{\boldsymbol{\theta}}^{(j)} + \frac{\hat{I}^{(j)-1}}{n}\left[\frac{\partial \log g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|_{\hat{\boldsymbol{\theta}}^{(j)}} - \frac{1}{2}\hat{H}^{(j)}\hat{I}^{(j)-1}\right] + O(n^{-2})$$

where $\tilde{\pi}_j = \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{X}_{[j]})$, $\hat{I}^{(j)} = -\frac{1}{m}\frac{\partial^2 \log f(X_{[j]}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(j)}}$, $\hat{H}^{(j)} = -\frac{1}{m}\frac{\partial^3 log f(X_{[j]}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T \partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(j)}}$, and $\hat{H}^{(j)}\hat{I}^{(j)-1}$ is a vector whose $r$th element equals $\sum_{st}\hat{H}^{(j)}_{rst}\hat{I}^{(j)-1}_{st}$. To simplify the notation, we denote $\hat{I}^{(j)-1}[\partial \log \pi(\boldsymbol{\theta})/\partial \boldsymbol{\theta}|_{\hat{\boldsymbol{\theta}}^{(j)}} - \frac{1}{2}\hat{H}^{(j)}\hat{I}^{(j)-1}]$ by $\boldsymbol{\nu}^{(j)}$. Moreover, for each $\hat{\boldsymbol{\theta}}^{(j)}$, we have

$$\hat{\boldsymbol{\theta}}^{(j)} = \boldsymbol{\theta}^* + \frac{\boldsymbol{\xi}^{(j)}}{\sqrt{m}} + O_p(m^{-1}),$$

where

$$\boldsymbol{\xi}^{(j)} = \frac{1}{\sqrt{m}} I^{-1} \sum_{i=1}^{m} \frac{\partial \log f(X_{ji}|\boldsymbol{\theta}^{(*)})}{\partial \boldsymbol{\theta}}, \quad I = -E_{X|\boldsymbol{\theta}^*}\frac{\partial^2 \log f(X|\boldsymbol{\theta}^{(*)})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}.$$

Therefore, the mean of the mixture distribution $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{X})$ is

$$E_{\tilde{\pi}}(\boldsymbol{\theta}) = \frac{1}{k}\sum_{j=1}^{k}\boldsymbol{\mu}^{(j)} = \boldsymbol{\theta}^* + \frac{1}{k}\sum_{j=1}^{k}\frac{\boldsymbol{\nu}^{(j)}}{n} + \frac{1}{k}\sum_{j=1}^{k}\frac{\boldsymbol{\xi}^{(j)}}{\sqrt{m}} + O_p(m^{-1}) + O(n^{-2}).$$

Note that $\boldsymbol{\nu}^{(j)}$ is of $O(1)$, $\boldsymbol{\xi}^{(j)}$'s are independent of each other and each has the mean 0 and variance $I^{-1}$. By condition (A3), we have

$$E[E_{\tilde{\pi}}(\boldsymbol{\theta}) - \boldsymbol{\theta}^{(*)}]^2 = \frac{k}{k^2}\frac{1}{m}diag(I^{-1}) + o(n^{-1}) = \frac{1}{n}diag(I^{-1}) + o(n^{-1}).$$

The variance of each subposterior can be approximated as follows:

$$Var_{\tilde{\pi}}(\boldsymbol{\theta}) = \frac{\hat{I}^{(j)-1}}{n} + O(n^{-2})$$

Therefore, the variance of the mixture distribution $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{X})$ is

$$Var_{\tilde{\pi}}(\boldsymbol{\theta}) = \frac{1}{k}\sum_{j=1}^{k} Var_{\tilde{\pi}}(\boldsymbol{\theta}) = \frac{1}{k}\sum_{j=1}^{k}\frac{\hat{I}^{(j)-1}}{n} + O(n^{-2}),$$

and

$$E[Var_{\tilde{\pi}}(\boldsymbol{\theta})] = \frac{1}{n}I^{-1} + o(n^{-1}).$$

By the definition of Wasserstein distance, we have

$$d^2(\tilde{\pi}, \delta_{\boldsymbol{\theta}^*}) = \int_{\Theta} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{X}) d\boldsymbol{\theta} = \|E_{\tilde{\pi}}(\boldsymbol{\theta}) - \boldsymbol{\theta}^*\|_2^2 + tr\left(Var_{\tilde{\pi}}(\boldsymbol{\theta})\right).$$

Then, from the above analysis, it is easy to see that

$$E(d^2(\tilde{\pi}, \delta_{\boldsymbol{\theta}^*})) = 2\frac{tr(I^{-1})}{n} + o(n^{-1}).$$

Following the same procedure, we can expand the full data posterior $\pi(\boldsymbol{\theta}|\boldsymbol{X})$:

$$\begin{aligned} E_\pi(\boldsymbol{\theta}) &= \boldsymbol{\theta}^{(*)} + \frac{\boldsymbol{\nu}}{n} + \frac{\boldsymbol{\xi}}{\sqrt{n}} + O_p(n^{-1}) + O(n^{-2}) \\ Var_\pi \boldsymbol{\theta} &= \frac{\hat{I}^{-1}}{n} + O(n^{-2}) \end{aligned}$$

where $\boldsymbol{\nu}$, $\boldsymbol{\xi}$ and $\hat{I}$ are defined correspondingly. Notice $\frac{\boldsymbol{\xi}}{\sqrt{n}} = \frac{1}{k}\sum_{j=1}^{k} \frac{\boldsymbol{\xi}^{(j)}}{\sqrt{m}}$, we thus have

$$\begin{aligned} E[E_{\tilde{\pi}}(\boldsymbol{\theta}) - E_\pi(\boldsymbol{\theta})]^2 &= O(m^{-2}), \\ E|Var_{\tilde{\pi}}(\boldsymbol{\theta}) - Var_\pi(\boldsymbol{\theta})| &= o(n^{-1}), \\ E(d^2(\pi, \delta_{\boldsymbol{\theta}^*})) &= 2\frac{tr(I^{-1})}{n} + o(n^{-1}), \end{aligned}$$

which completes the proof.

APPENDIX C

SUPPLEMENTARY MATERIALS FOR CHAPTER 4

In the following text, we use $l(D_n|\boldsymbol{\beta}_s)$ to denote the log-likelihood $\log L(D_n|\boldsymbol{\beta}_s)$, use $\mathbf{s}(D_n, \boldsymbol{\beta}_s)$ to denote the score function $\partial l(D_n|\boldsymbol{\beta}_s)/\partial\boldsymbol{\beta}_s$, and use $H(D_n, \boldsymbol{\beta}_s)$ to denote the observed information matrix $-\partial^2 l(D_n|\boldsymbol{\beta}_s)/\partial\boldsymbol{\beta}_s\partial\boldsymbol{\beta}_s^T$. So $H(D_n, \boldsymbol{\beta}_s) = n\widehat{I}(D_n|\boldsymbol{\beta}_s)$. Moreover, we introduce $\boldsymbol{\beta}_s^*$(for $s \supseteq s^*$) to denote the sub-vector of $\boldsymbol{\beta}^*$ corresponding to the model $s$. For simplicity, we also depress the subscript of $p_n$ and replace it by $p$.

Before giving the proof of Theorem 4.1 and Theorem 4.2, we first present a useful lemma, which is modified from Lemma1 in Foygel and Drton (2011).

## C.1 A Useful Lemma

**Lemma C.1.** *Assume the conditions (A1)-(A5) are satisfied, then for sufficiently large $n$, with probability at least $1 - n^{-\eta}$ ($\eta$ is any positive number) , the following statements all hold.*

1. *For all $|s| \leq 2q$, $||\boldsymbol{\beta}_s||_2 \leq r$ and $||\boldsymbol{\beta}'_s||_2 \leq r$, where $r > 0$, There exists positive numbers $b_1(r)$,$b_2(r)$ and $b_3(r)$ such that*

$$b_1(r) \leq \lambda_{\min}[\frac{1}{n}H(D_n, \boldsymbol{\beta}_s)] \leq \lambda_{\max}[\frac{1}{n}H(D_n, \boldsymbol{\beta}_s)] \leq b_2(r)$$

$$-b_3(r)||\boldsymbol{\beta}_s - \boldsymbol{\beta'}_s||_2 \leq \lambda_{\min}[\frac{1}{n}H(D_n, \boldsymbol{\beta}_s) - \frac{1}{n}H(D_n, \boldsymbol{\beta'}_s)]$$

$$b_3(r)||\boldsymbol{\beta}_s - \boldsymbol{\beta'}_s||_2 \geq \lambda_{\max}[\frac{1}{n}H(D_n, \boldsymbol{\beta}_s) - \frac{1}{n}H(D_n, \boldsymbol{\beta'}_s)]$$

2. $||H(D_n, \boldsymbol{\beta}_s^*)^{-1/2}\mathbf{s}(D_n, \boldsymbol{\beta}_s^*)||_2 \leq \sqrt{2(1+\epsilon_n)|s\backslash s^*|\log(n^\eta p)}$ *for all $s \supseteq s^*$ with*

$|s| \leq 2q$, *where* $\epsilon_n = \log^{-1/2}(n) = o(1)$. *In the case* $s = s^*$, *we define* $|s \backslash s^*|$ *as 1.*

3. $l(D_n|\boldsymbol{\beta}_s^* + \boldsymbol{\psi}_s) - l(D_n|\boldsymbol{\beta}_s^*) \leq -\frac{b_1(M+1)n}{2}||\boldsymbol{\psi}_s||_2 \left[ \min(1, ||\boldsymbol{\psi}_s||_2) - \tau\sqrt{\frac{\log(n^\eta p)}{n}} \right]$ *for all* $s \supseteq s^*$ *with* $|s| \leq 2q$ *and* $\boldsymbol{\psi}_s \in \mathbb{R}^{|s|}$, *where* $\tau = \sqrt{\frac{32qb_2(M+1)}{b_1^2(M+1)}}$

4. $||\widehat{\boldsymbol{\beta}}_s||_2 \leq R$ *for all model* $s$ *with* $|s| \leq q$ *and* $s \not\supseteq s^*$, *where* $R = M + 1 + \frac{4b_2(M+1)M^2}{b_1(M+1)}$

*Proof.* In the generalized linear model, we have

$$\begin{aligned} l(D_n|\boldsymbol{\beta}_s) &= \sum_{i=1}^{n}[y^{(i)}(\mathbf{x}_s^{(i)})^T\boldsymbol{\beta}_s - b((\mathbf{x}_s^{(i)})^T\boldsymbol{\beta}_s)] \\ \mathbf{s}(D_n, \boldsymbol{\beta}_s) &= \sum_{i=1}^{n}[y^{(i)} - b'((\mathbf{x}_s^{(i)})^T\boldsymbol{\beta}_s)]\mathbf{x}_s^{(i)} \\ H(D_n, \boldsymbol{\beta}_s) &= \sum_{i=1}^{n} b''((\mathbf{x}_s^{(i)})^T\boldsymbol{\beta}_s)\mathbf{x}_s^{(i)}(\mathbf{x}_s^{(i)})^T \end{aligned}$$

Notice $H(D_n, \boldsymbol{\beta}_s)$ does not depend on the response variable $y$

**Part 1**

$|x_j^{(i)}| \leq 1$ and $||\boldsymbol{\beta}_s||_2 \leq r$ lead to $|(\mathbf{x}_s^{(i)})^T\boldsymbol{\beta}_s| \leq 2rq$. Define $d_1(r) = \inf_{|\theta|\leq 2rq} b''(\theta)$ and $d_2(r) = \sup_{|\theta|\leq 2rq} b''(\theta)$. We have

$$\lambda_{\min}(\frac{1}{n}H(D_n, \boldsymbol{\beta}_s)) \geq \min_i(b''((\mathbf{x}_s^{(i)})^T\boldsymbol{\beta}_s))\lambda_{\min}(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_s^{(i)}(\mathbf{x}_s^{(i)})^T) \geq d_1(r)a_1$$

$$\lambda_{\max}(\frac{1}{n}H(D_n, \boldsymbol{\beta}_s)) \leq \max_i(b''((\mathbf{x}_s^{(i)})^T\boldsymbol{\beta}_s))\lambda_{\min}(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_s^{(i)}(\mathbf{x}_s^{(i)})^T) \leq d_2(r)a_2$$

In addition

$$\begin{aligned} H(D_n, \boldsymbol{\beta}_s) - H(D_n, \boldsymbol{\beta'}_s) &= \sum_{i=1}^{n} \mathbf{x}_s^{(i)} (\mathbf{x}_s^{(i)})^T [b''((\mathbf{x}_s^{(i)})^T \boldsymbol{\beta}_s) - b''((\mathbf{x}_s^{(i)})^T \boldsymbol{\beta'}_s)] \\ &= \sum_{i=1}^{n} \mathbf{x}_s^{(i)} (\mathbf{x}_s^{(i)})^T b'''((\mathbf{x}_s^{(i)})^T \boldsymbol{\beta''}_s)[(\mathbf{x}_s^{(i)})^T (\boldsymbol{\beta}_s - \boldsymbol{\beta'}_s)] \end{aligned}$$

where $\boldsymbol{\beta''}_s$ is between $\boldsymbol{\beta}_s$ and $\boldsymbol{\beta'}_s$. Define $d_3(r) = \sup_{|\theta| \leq 2rq} b'''(\theta)$ and notice $|(\mathbf{x}_s^{(i)})^T (\boldsymbol{\beta}_s - \boldsymbol{\beta'}_s)| \leq 2q||\boldsymbol{\beta}_s - \boldsymbol{\beta'}_s||_2$, we further have

$$\begin{aligned} \lambda_{\max}[H(D_n, \boldsymbol{\beta}_s) - H(D_n, \boldsymbol{\beta'}_s)] &\leq \lambda_{\max}\{\sum_{i=1}^{n} \mathbf{x}_s^{(i)} (\mathbf{x}_s^{(i)})^T\} d_3(r) 2q||\boldsymbol{\beta}_s - \boldsymbol{\beta'}_s||_2 \\ &\leq 2qa_2 d_3(r)||\boldsymbol{\beta}_s - \boldsymbol{\beta'}_s||_2 \\ \lambda_{\min}[H(D_n, \boldsymbol{\beta}_s) - H(D_n, \boldsymbol{\beta'}_s)] &\geq \lambda_{\max}\{\sum_{i=1}^{n} \mathbf{x}_s^{(i)} (\mathbf{x}_s^{(i)})^T\} (-1) d_3(r) 2q||\boldsymbol{\beta}_s - \boldsymbol{\beta'}_s||_2 \\ &\geq -2qa_2 d_3(r)||\boldsymbol{\beta}_s - \boldsymbol{\beta'}_s||_2 \end{aligned}$$

**Part 2**

For any model $s$ with $s \supseteq s^*, |s| \leq 2q$ and any vector $\mathbf{u} \in \mathbb{R}^{|s|}$ with $||\mathbf{u}||_2 \leq 1$, we have

$$\begin{aligned} \mathbf{u}^T H(D_n, \boldsymbol{\beta}_s^*)^{-1/2} \mathbf{s}(D_n, \boldsymbol{\beta}_s^*) &= \sum_{i=1}^{n} [y^{(i)} - b'((\mathbf{x}_s^{(i)})^T \boldsymbol{\beta}_s^*)](\mathbf{x}_s^{(i)})^T H(D_n, \boldsymbol{\beta}_s^*)^{-1/2} \mathbf{u} \\ &= \sum_{i=1}^{n} [y^{(i)} - \mu^{(i)}](\mathbf{x}_s^{(i)})^T H(D_n, \boldsymbol{\beta}_s^*)^{-1/2} \mathbf{u} \end{aligned}$$

Now define $A_{s,n} = \sqrt{2\sqrt{1+\epsilon_n}|s \backslash s^*| \log(n^\eta p)}$ and $\boldsymbol{\psi}_s = A_{s,n} H(D_n, \boldsymbol{\beta}_s^*)^{-1/2} \mathbf{u}$, we have

$$
\begin{aligned}
& P\{\mathbf{u}^T H(D_n, \boldsymbol{\beta}_s^*)^{-1/2} \mathbf{s}(D_n, \boldsymbol{\beta}_s^*) \geq A_{s,n}\} \\
= \; & E\Big[\mathbb{1}\{A_{s,n}\mathbf{u}^T H(D_n, \boldsymbol{\beta}_s^*)^{-1/2} \mathbf{s}(D_n, \boldsymbol{\beta}_s^*) \geq A_{s,n}^2\}\Big] \\
= \; & E\Big[\mathbb{1}\{\sum_{i=1}^{n} [y^{(i)} - \mu^{(i)}](\mathbf{x}_s^{(i)})^T \boldsymbol{\psi}_s \geq A_{s,n}^2\}\Big] \\
\leq \; & E\Big[\exp\{\sum_{i=1}^{n} [y^{(i)} - \mu^{(i)}](\mathbf{x}_s^{(i)})^T \boldsymbol{\psi}_s - A_{s,n}^2\}\Big] \\
= \; & \exp\{\sum_{i=1}^{n} -\mu^{(i)}(\mathbf{x}_s^{(i)})^T \boldsymbol{\psi}_s - A_{s,n}^2\} E\Big[\sum_{i=1}^{n} \exp\{y^{(i)}(\mathbf{x}_s^{(i)})^T \boldsymbol{\psi}_s\}\Big] \\
= \; & \exp\{\sum_{i=1}^{n} -\mu^{(i)}(\mathbf{x}_s^{(i)})^T \boldsymbol{\psi}_s - A_{s,n}^2\} \prod_{i=1}^{n} E\Big[\exp\{y^{(i)}(\mathbf{x}_s^{(i)})^T \boldsymbol{\psi}_s\}\Big] \\
= \; & \exp\{\sum_{i=1}^{n} -\mu^{(i)}(\mathbf{x}_s^{(i)})^T \boldsymbol{\psi}_s - A_{s,n}^2\} \prod_{i=1}^{n} \exp\{b((\mathbf{x}_s^{(i)})^T(\boldsymbol{\psi}_s + \boldsymbol{\beta}_s^*)) - b((\mathbf{x}_s^{(i)})^T \boldsymbol{\beta}_s^*)\} \\
= \; & \exp\{\sum_{i=1}^{n} -\mu^{(i)}(\mathbf{x}_s^{(i)})^T \boldsymbol{\psi}_s - A_{s,n}^2\} \exp\{\sum_{i=1}^{n} [b((\mathbf{x}_s^{(i)})^T(\boldsymbol{\psi}_s + \boldsymbol{\beta}_s^*)) - b((\mathbf{x}_s^{(i)})^T \boldsymbol{\beta}_s^*)]\}
\end{aligned}
$$

By Taylor expansion, we can rewrite the contents in the second exponential function

$$
\begin{aligned}
& \sum_{i=1}^{n} [b((\mathbf{x}_s^{(i)})^T(\boldsymbol{\psi}_s + \boldsymbol{\beta}_s^*)) - b((\mathbf{x}_s^{(i)})^T \boldsymbol{\beta}_s^*)] \\
= \; & \sum_{i=1}^{n} \Big\{ b'((\mathbf{x}_s^{(i)})^T \boldsymbol{\beta}_s^*)(\mathbf{x}_s^{(i)})^T \psi(D_n) + \frac{1}{2} b''((\mathbf{x}_s^{(i)})^T \boldsymbol{\beta}_s^*)[(\mathbf{x}_s^{(i)})^T \psi(D_n)]^2 \\
& + \frac{1}{2}[b''((\mathbf{x}_s^{(i)})^T(\boldsymbol{\psi}_s + \boldsymbol{\beta}_s^*)) - b''((\mathbf{x}_s^{(i)})^T \boldsymbol{\beta}_s^*)][(\mathbf{x}_s^{(i)})^T \boldsymbol{\psi}_s]^2 \Big\} \\
= \; & \sum_{i=1}^{n} \mu^{(i)}(\mathbf{x}_s^{(i)})^T \boldsymbol{\psi}_s + \sum_{i=1}^{n} \frac{1}{2} b''((\mathbf{x}_s^{(i)})^T \boldsymbol{\beta}_s^*)[(\mathbf{x}_s^{(i)})^T \boldsymbol{\psi}_s]^2 \\
& + \frac{1}{2} \boldsymbol{\psi}_s^T \sum_{i=1}^{n} [b''((\mathbf{x}_s^{(i)})^T(\boldsymbol{\psi}_s + \boldsymbol{\beta}_s^*)) - b''((\mathbf{x}_s^{(i)})^T \boldsymbol{\beta}_s^*)][\mathbf{x}_s^{(i)}(\mathbf{x}_s^{(i)})^T] \boldsymbol{\psi}_s
\end{aligned}
$$

To step it further, we have to analyze the second term and third term. For the

second one, notice

$$\begin{aligned}
&\sum_{i=1}^{n} b''((\mathbf{x}_s^{(i)})^T\boldsymbol{\beta}_s^*)[(\mathbf{x}_s^{(i)})^T\boldsymbol{\psi}_s]^2 \\
= \;& A_{s,n}^2 \sum_{i=1}^{n} b''((\mathbf{x}_s^{(i)})^T\boldsymbol{\beta}_s^*)[(\mathbf{x}_s^{(i)})^T H(D_n,\boldsymbol{\beta}_s^*)^{-1/2}\mathbf{u}]^2 \\
= \;& A_{s,n}^2 \mathbf{u}^T H(D_n,\boldsymbol{\beta}_s^*)^{-1/2}\left[\sum_{i=1}^{n}(\mathbf{x}_s^{(i)})^T b''((\mathbf{x}_s^{(i)})^T\boldsymbol{\beta}_s^*)\mathbf{x}_s^{(i)}\right] H(D_n,\boldsymbol{\beta}_s^*)^{-1/2}\mathbf{u} \\
= \;& A_{s,n}^2 \mathbf{u}^T H(D_n,\boldsymbol{\beta}_s^*)^{-1/2} H(D_n,\boldsymbol{\beta}_s^*) H(D_n,\boldsymbol{\beta}_s^*)^{-1/2}\mathbf{u} \\
= \;& A_{s,n}^2 \mathbf{u}^T\mathbf{u} \leq A_{s,n}^2
\end{aligned}$$

For the third one, we can apply part 1 of the lemma, with $r$ set as $M+1$, and get

$$||\boldsymbol{\psi}_s||_2^2 = A_{s,n}^2 \mathbf{u}^T H(D_n,\boldsymbol{\beta}_s^*)^{-1}\mathbf{u} \leq A_{s,n}^2 \frac{1}{nb_1(M+1)}\mathbf{u}^T\mathbf{u} \leq \frac{A_{s,n}^2}{nb_1(M+1)}$$

We should also observe that for $n$ sufficiently large,

$$||\boldsymbol{\psi}_s||_2^2 \leq \frac{A_{s,n}^2}{nb_1(M+1)} = \frac{2\sqrt{1+\epsilon_n}|s\backslash s^*|\log(n^\eta p)}{nb_1(M+1)} \leq \frac{5q\log(n^\eta p)}{nb_1(M+1)} < 1$$

Then apply part 1 of the lemma again, with $r$ set as $M+1$

$$\begin{aligned}
&\boldsymbol{\psi}_s^T \sum_{i=1}^{n}[b''((\mathbf{x}_s^{(i)})^T(\boldsymbol{\psi}_s+\boldsymbol{\beta}_s^*)) - b''((\mathbf{x}_s^{(i)})^T\boldsymbol{\beta}_s^*)][\mathbf{x}_s^{(i)}(\mathbf{x}_s^{(i)})^T]\boldsymbol{\psi}_s \\
= \;& \boldsymbol{\psi}_s^T[H(D_n,\boldsymbol{\psi}_s+\boldsymbol{\beta}_s^*) - H(D_n,\boldsymbol{\beta}_s^*)]\boldsymbol{\psi}_s \\
\leq \;& nb_3(M+1)||\boldsymbol{\psi}_s||_2\boldsymbol{\psi}_s^T\boldsymbol{\psi}_s = \frac{1}{2}nb_3(M+1)||\boldsymbol{\psi}_s||_2^3 \\
\leq \;& nb_3(M+1)\frac{A_{s,n}^3}{n^{1.5}b_1^{1.5}(M+1)} = \frac{A_{s,n}^3 b_3(M+1)}{n^{0.5}b_1^{1.5}(M+1)}
\end{aligned}$$

By combining the above inequalities together, we finally have

$$\begin{aligned} & \sum_{i=1}^{n}[b((\mathbf{x}_s^{(i)})^T(\boldsymbol{\psi}_s+\boldsymbol{\beta}_s^*))-b((\mathbf{x}_s^{(i)})^T\boldsymbol{\beta}_s^*)] \\ \leq & \sum_{i=1}^{n}\mu^{(i)}(\mathbf{x}_s^{(i)})^T\boldsymbol{\psi}_s+\frac{A_{s,n}^2}{2}+\frac{A_{s,n}^3 b_3(M+1)}{2n^{0.5}b_1^{1.5}(M+1)} \end{aligned}$$

Now go back to the analysis of $P\{\mathbf{u}^T H(D_n,\boldsymbol{\beta}_s^*)^{-1/2}\mathbf{s}(D_n,\boldsymbol{\beta}_s^*)\geq A_{s,n}\}$ and continue,

$$\begin{aligned} & P\{\mathbf{u}^T H(D_n,\boldsymbol{\beta}_s^*)^{-1/2}\mathbf{s}(D_n,\boldsymbol{\beta}_s^*)\geq A_{s,n}\} \\ \leq & \exp\{\sum_{i=1}^{n}-\mu^{(i)}(\mathbf{x}_s^{(i)})^T\boldsymbol{\psi}_s-A_{s,n}^2\}\exp\{\sum_{i=1}^{n}[b((\mathbf{x}_s^{(i)})^T(\boldsymbol{\psi}_s+\boldsymbol{\beta}_s^*))-b((\mathbf{x}_s^{(i)})^T\boldsymbol{\beta}_s^*)]\} \\ \leq & \exp\{\sum_{i=1}^{n}-\mu^{(i)}(\mathbf{x}_s^{(i)})^T\boldsymbol{\psi}_s-A_{s,n}^2\}\exp\{\sum_{i=1}^{n}\mu^{(i)}(\mathbf{x}_s^{(i)})^T\boldsymbol{\psi}_s+\frac{A_{s,n}^2}{2}+\frac{A_{s,n}^3 b_3(M+1)}{2n^{0.5}b_1^{1.5}(M+1)}\} \\ = & \exp\{-\frac{A_{s,n}^2}{2}+\frac{A_{s,n}^3 b_3(M+1)}{2n^{0.5}b_1^{1.5}(M+1)}\} \end{aligned}$$

To link this inequality to our final objective, we rely on lemma 2 in Chen and Chen (2012), which states that for a given $\delta_n>0$, there exists a finite set of unit vectors $\mathbf{U}(\delta_n)\subset\mathbb{R}^{2q}$ such that for all $\mathbf{v}\in\mathbb{R}^{2q}$, we have $||\mathbf{v}||_2\leq(1+\delta_n)\max_{\mathbf{u}\in\mathbf{U}(\delta_n)}\mathbf{u}^T\mathbf{v}$. Since $\mathbf{U}(\delta_n)$ is a finite set, we use $N(\delta_n)$ to denote its cardinality.

In fact, in the following, we use a corollary of this lemma, that is, for $\mathbf{v}$ with length $l\leq 2q$, we have

$$||\mathbf{v}||_2\leq(1+\delta_n)\max_{\mathbf{u}\in\mathbf{U}(\delta_n)}\mathbf{u}_l^T\mathbf{v}$$

where $\mathbf{u}_l$ denotes the first $l$ elements in $\mathbf{u}$. This corollary can be verified easily.

So for a fixed model $s$ with $s\supseteq s^*$ and $|s|=d\leq 2q$. If we let $\delta_n=\sqrt[4]{1+\epsilon_n}-1$,

then we have

$$
\begin{aligned}
& P\{||H(D_n,\boldsymbol{\beta}_s^*)^{-1/2}\mathbf{s}(D_n,\boldsymbol{\beta}_s^*)||_2 \geq (1+\delta_n)A_{s,n}\} \\
\leq & \sum_{\mathbf{u}\in\mathbf{U}(\delta_n)} P\{\mathbf{u}_d^T H(D_n,\boldsymbol{\beta}_s^*)^{-1/2}\mathbf{s}(D_n,\boldsymbol{\beta}_s^*) \geq A_{s,n}\} \\
\leq & \sum_{\mathbf{u}\in\mathbf{U}(\delta_n)} \exp\{-\frac{A_{s,n}^2}{2}+\frac{A_{s,n}^3 b_3(M+1)}{2n^{0.5}b_1^{1.5}(M+1)}\} \\
= & N(\delta_n)\exp\{-\frac{A_{s,n}^2}{2}+\frac{A_{s,n}^3 b_3(M+1)}{2n^{0.5}b_1^{1.5}(M+1)}\}
\end{aligned}
$$

For neatness and conciseness, we only give the proof excluding $s = s^*$

$$
\begin{aligned}
& P\{\exists\ s \text{ with } s\supset s^* \text{ and } |s|\leq 2q, ||H(D_n,\boldsymbol{\beta}_s^*)^{-1/2}\mathbf{s}(D_n,\boldsymbol{\beta}_s^*)||_2 \geq (1+\delta_n)A_{s,n}\} \\
\leq & \sum_{s\supset s^* \text{ and } |s|\leq 2q} P\{||H(D_n,\boldsymbol{\beta}_s^*)^{-1/2}\mathbf{s}(D_n,\boldsymbol{\beta}_s^*)||_2 \geq (1+\delta_n)A_{s,n}\} \\
\leq & \sum_{s\supset s^* \text{ and } |s|\leq 2q} N(\delta_n)\exp\{-\frac{A_{s,n}^2}{2}+\frac{A_{s,n}^3 b_3(M+1)}{2n^{0.5}b_1^{1.5}(M+1)}\} \\
= & \sum_{d'=1}^{2q-|s^*|}\binom{p}{d'} N(\delta_n)\exp\{-\frac{A_{s,n}^2}{2}+\frac{A_{s,n}^3 b_3(M+1)}{2n^{0.5}b_1^{1.5}(M+1)}\} \\
\leq & \sum_{d'=1}^{2q-|s^*|}\exp\{-\sqrt{1+\epsilon_n}d'\log(n^\eta p)\left[1-\sqrt{\frac{2(1+\epsilon_n)d'\log(n^\eta p)b_3^2(M+1)}{nb_1^3(M+1)}}\right] \\
& +d'\log(p)+\log(N(\delta_n))\}
\end{aligned}
$$

Now we analyze the term inside exponential function by splitting it into the following two component

$$
\begin{aligned}
& -\sqrt{1+\epsilon_n}d'\log(p)\left[1-\sqrt{\frac{2(1+\epsilon_n)d'\log(n^\eta p)b_3^2(M+1)}{nb_1^3(M+1)}}\right]+d'\log(p) \\
= & d'\log(p)\left\{1-\sqrt{1+\epsilon_n}\left[1-\sqrt{\frac{2(1+\epsilon_n)d'\log(n^\eta p)b_3^2(M+1)}{nb_1^3(M+1)}}\right]\right\}
\end{aligned}
$$

and

$$
\begin{aligned}
& -\sqrt{1+\epsilon_n} d' \log(n^\eta) \left[ 1 - \sqrt{\frac{2(1+\epsilon_n) d' \log(n^\eta p) b_3^2 (M+1)}{n b_1^3 (M+1)}} \right] + \log(N(\delta_n)) \\
= \quad & d' \log(n^\eta) \left\{ -\sqrt{1+\epsilon_n} \left[ 1 - \sqrt{\frac{2(1+\epsilon_n) d' \log(n^\eta p) b_3^2 (M+1)}{n b_1^3 (M+1)}} \right] + \frac{\log(N(\delta_n))}{d' \log(n^\eta)} \right\}
\end{aligned}
$$

For the first component, we should notice when the $n$ is sufficiently large

$$
\sqrt{\frac{2(1+\epsilon_n) d' \log(n^\eta p) b_3^2 (M+1)}{n b_1^3 (M+1)}} = O(\sqrt{\frac{\log(n^\eta p)}{n}}) = o(\frac{1}{\log^{1/2}(n)}) = o(\epsilon_n)
$$

So $\sqrt{1+\epsilon_n}[1 - \sqrt{\frac{2(1+\epsilon_n) d' \log(n^\eta p) b_3^2 (M+1)}{n b_1^3 (M+1)}}] = 1 + O(\epsilon_n)$ should be larger than 1 for $n$ sufficiently large, and component 1 should be smaller than 0.

For the second component, we should further notice when $n$ is sufficiently large

$$
N(\delta_n) = O(1/\delta_n) = O(\frac{1}{\sqrt[4]{1+\epsilon_n} - 1}) = O(\frac{1}{\epsilon_n}) = O(\log^{1/2}(n))
$$

so

$$
\frac{\log(N(\delta_n))}{d' \log(n^\eta)} = O(\frac{\log\log(n)}{\log(n)}) = o(\frac{1}{\log^{1/2}(n)}) = o(\epsilon_n)
$$

and component 2 becomes

$$
d' \log(n^\eta)(-1 - O(\epsilon_n)) = -d' \log(n^\eta) - O(\log^{1/2}(n))
$$

For $n$ sufficiently large, $O(\log^{1/2}(n)) > \log(4)$.

Combing them together, we finally have

$$\begin{aligned}
& P\{\exists\ s \text{ with } s \supset s^* \text{ and } |s| \le 2q, ||H(D_n, \boldsymbol{\beta}_s^*)^{-1/2}\mathbf{s}(D_n, \boldsymbol{\beta}_s^*)||_2 \ge (1+\delta_n)A_{s,n}\} \\
\le & \sum_{d'=1}^{2q-|s^*|} \exp\{-\sqrt{1+\epsilon_n}d' \log(n^\eta p)\left[1 - \sqrt{\frac{2(1+\epsilon_n)d' \log(n^\eta p) b_3^2(M+1)}{n b_1^3 (M+1)}}\right] \\
& + d' \log(p) + \log(N(\delta_n))\} \\
\le & \sum_{d'=1}^{2q-|s^*|} \exp\{-d' \log(n^\eta) - \log(4)\} < \sum_{d'=1}^{\infty} \frac{1}{4}\exp\{-d' \log(n^\eta)\} \le \frac{1}{2} n^{-\eta}
\end{aligned}$$

Following very similar procedure, we can also obtain

$$P\{\exists\ s \text{ with } s \supseteq s^* \text{ and } |s| \le 2q, ||H(D_n, \boldsymbol{\beta}_s^*)^{-1/2}\mathbf{s}(D_n, \boldsymbol{\beta}_s^*)||_2 \ge (1+\delta_n)A_{s,n}\} \le \frac{1}{2} n^{-\eta}$$

**Part 3**

Suppose part 2 of this lemma holds, then for any model $s$ with $s \supseteq s^*, |s| \le 2q$ and $\boldsymbol{\psi}_s$ with $||\boldsymbol{\psi}_s||_2 \le 1$, we have

$$\begin{aligned}
& l(D_n|\boldsymbol{\beta}_s^* + \boldsymbol{\psi}_s) - l(D_n|\boldsymbol{\beta}_s^*) \\
= & \boldsymbol{\psi}_s^T \mathbf{s}(D_n, \boldsymbol{\beta}_s^*) - \frac{1}{2}\boldsymbol{\psi}_s^T H(D_n, \boldsymbol{\beta}_s^* + t\boldsymbol{\psi}_s)\boldsymbol{\psi}_s \\
\le & ||\boldsymbol{\psi}_s||_2 ||\mathbf{s}(D_n, \boldsymbol{\beta}_s^*)||_2 - \frac{b_1(M+1)n}{2}||\boldsymbol{\psi}_s||_2^2 \\
= & ||\boldsymbol{\psi}_s||_2 ||H(D_n, \boldsymbol{\beta}_s^*)^{1/2} H(D_n, \boldsymbol{\beta}_s^*)^{-1/2}\mathbf{s}(D_n, \boldsymbol{\beta}_s^*)||_2 - \frac{b_1(M+1)n}{2}||\boldsymbol{\psi}_s||_2^2 \\
\le & ||\boldsymbol{\psi}_s||_2 \sqrt{\lambda_{\max}(H(D_n, \boldsymbol{\beta}_s^*))}||H(D_n, \boldsymbol{\beta}_s^*)^{-1/2}\mathbf{s}(D_n, \boldsymbol{\beta}_s^*)||_2 - \frac{b_1(M+1)n}{2}||\boldsymbol{\psi}_s||_2^2 \\
\le & ||\boldsymbol{\psi}_s||_2 \sqrt{n b_2 (M+1)}\sqrt{2(1+\epsilon_n)|s \backslash s^*| \log(n^\eta p)} - \frac{b_1(M+1)n}{2}||\boldsymbol{\psi}_s||_2^2 \\
\le & -\frac{b_1(M+1)n}{2}||\boldsymbol{\psi}_s||_2 \left[||\boldsymbol{\psi}_s||_2 - \sqrt{\frac{\log(n^\eta p)}{n}}\sqrt{\frac{32 q b_2 (M+1)}{b_1^2 (M+1)}}\right] \\
\le & -\frac{b_1(M+1)n}{2}||\boldsymbol{\psi}_s||_2 \left[||\boldsymbol{\psi}_s||_2 - \tau\sqrt{\frac{\log(n^\eta p)}{n}}\right]
\end{aligned}$$

From the above inequality, we also obtain $||\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_s^*||_2 \leq \tau\sqrt{\frac{\log(n^\eta p)}{n}} = o(1)$, which means for $n$ sufficiently large, $\widehat{\boldsymbol{\beta}}_s$ and $\boldsymbol{\beta}_s^*$ are very close with each other. Next, by concavity of the log-likelihood, for $||\boldsymbol{\psi}_s||_2 > 1$, we have

$$l(D_n|\widehat{\boldsymbol{\beta}}_s + \boldsymbol{\psi}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s) \leq ||\boldsymbol{\psi}_s||_2[l(D_n|\widehat{\boldsymbol{\beta}}_s + \frac{\boldsymbol{\psi}_s}{||\boldsymbol{\psi}_s||_2}) - l(D_n|\widehat{\boldsymbol{\beta}}_s)]$$

By replacing $\widehat{\boldsymbol{\beta}}_s$ by $\boldsymbol{\beta}_s^*$ , now we have

$$\begin{aligned} l(D_n|\boldsymbol{\beta}_s^* + \boldsymbol{\psi}_s) - l(D_n|\boldsymbol{\beta}_s^*) &\leq ||\boldsymbol{\psi}_s||_2[l(D_n|\boldsymbol{\beta}_s^* + \frac{\boldsymbol{\psi}_s}{||\boldsymbol{\psi}_s||_2}) - l(D_n|\boldsymbol{\beta}_s^*)] \\ &\leq ||\boldsymbol{\psi}_s||_2\bigg[-\frac{b_1(M+1)n}{2}\bigg(1 - \tau\sqrt{\frac{\log(n^\eta p)}{n}}\bigg)\bigg] \\ &= -\frac{b_1(M+1)n}{2}||\boldsymbol{\psi}_s||_2\bigg[1 - \tau\sqrt{\frac{\log(n^\eta p)}{n}}\bigg] \end{aligned}$$

By combining these two cases together, we finally proved

$$l(D_n|\boldsymbol{\beta}_s^* + \boldsymbol{\psi}_s) - l(D_n|\boldsymbol{\beta}_s^*) \leq -\frac{b_1(M+1)n}{2}||\boldsymbol{\psi}_s||_2\bigg[\min(1, ||\boldsymbol{\psi}_s||_2) - \tau\sqrt{\frac{\log(n^\eta p)}{n}}\bigg]$$

**Part 4**

Suppose part 2 of this lemma holds (part 3 also holds). For any model $s$ with $|s| \leq q$ and $s \not\supseteq s^*$, let $s' = s \cup s^*$. It is easy to verify $s' \subseteq s^*$ and $|s'| \leq 2q$. We also use $\hat{\boldsymbol{\beta}}_{s,s'}$ to denote a vector corresponding to model $s'$, generated by $\hat{\boldsymbol{\beta}}_s$ augmented with zeros in $s' \backslash s$. Recall $R = M + 1 + \frac{4b_2(M+1)M^2}{b_1(M+1)}$, if $||\widehat{\boldsymbol{\beta}}_{s,s'} - \boldsymbol{\beta}_{s'}^*||_2 \leq 1$ then $||\widehat{\boldsymbol{\beta}}_{s,s'}||_2 \leq R$ obviously. So we only consider the case where $||\widehat{\boldsymbol{\beta}}_{s,s'} - \boldsymbol{\beta}_{s'}^*||_2 > 1$

By part 3 of this lemma, for $n$ sufficiently large, we have

$$\begin{aligned}
& l(D_n|\widehat{\boldsymbol{\beta}}_{s,s'}) - l(D_n|\boldsymbol{\beta}^*_{s'}) \\
\leq\ & -\frac{b_1(M+1)n}{2}||\widehat{\boldsymbol{\beta}}_{s,s'} - \boldsymbol{\beta}^*_{s'}||_2\left[\min(1, ||\widehat{\boldsymbol{\beta}}_{s,s'} - \boldsymbol{\beta}^*_{s'}||_2) - \tau\sqrt{\frac{\log(n^\eta p)}{n}}\right] \\
\leq\ & -\frac{b_1(M+1)n}{2}||\widehat{\boldsymbol{\beta}}_{s,s'} - \boldsymbol{\beta}^*_{s'}||_2\left[1 - \tau\sqrt{\frac{\log(n^\eta p)}{n}}\right] \\
\leq\ & -\frac{b_1(M+1)n}{4}||\widehat{\boldsymbol{\beta}}_{s,s'} - \boldsymbol{\beta}^*_{s'}||_2
\end{aligned}$$

We also have

$$\begin{aligned}
& l(D_n|\mathbf{0}_{s'}) - l(D_n|\boldsymbol{\beta}^*_{s'}) \\
=\ & (-\boldsymbol{\beta}^*_{s'})^T\mathbf{s}(D_n, \boldsymbol{\beta}^*_{s'}) - \frac{1}{2}(-\boldsymbol{\beta}^*_{s'})^T H(D_n, t\boldsymbol{\beta}^*_{s'})(-\boldsymbol{\beta}^*_{s'}) \\
=\ & -(H(D_n, \boldsymbol{\beta}^*_{s'})^{1/2}\boldsymbol{\beta}^*_{s'})^T H(D_n, \boldsymbol{\beta}^*_{s'})^{-1/2}\mathbf{s}(D_n, \boldsymbol{\beta}^*_{s'}) - \frac{1}{2}(-\boldsymbol{\beta}^*_{s'})^T H(D_n, t\boldsymbol{\beta}^*_{s'})(-\boldsymbol{\beta}^*_{s'}) \\
\geq\ & -||H(D_n, \boldsymbol{\beta}^*_{s'})^{1/2}\boldsymbol{\beta}^*_{s'}||_2||H(D_n, \boldsymbol{\beta}^*_{s'})^{-1/2}\mathbf{s}(D_n, \boldsymbol{\beta}^*_{s'})||_2 - \frac{1}{2}(\boldsymbol{\beta}^*_{s'})^T H(D_n, t\boldsymbol{\beta}^*_{s'})(\boldsymbol{\beta}^*_{s'}) \\
=\ & -\sqrt{(\boldsymbol{\beta}^*_{s'})^T H(D_n, \boldsymbol{\beta}^*_{s'})\boldsymbol{\beta}^*_{s'}}||H(D_n, \boldsymbol{\beta}^*_{s'})^{-1/2}\mathbf{s}(D_n, \boldsymbol{\beta}^*_{s'})||_2 - \frac{1}{2}(\boldsymbol{\beta}^*_{s'})^T H(D_n, t\boldsymbol{\beta}^*_{s'})(\boldsymbol{\beta}^*_{s'}) \\
\geq\ & -\sqrt{nb_2(M+1)(\boldsymbol{\beta}^*_{s'})^T\boldsymbol{\beta}^*_{s'}}\sqrt{2(1+\epsilon_n)|s'\backslash s^*|\log(n^\eta p)} - \frac{1}{2}nb_2(M+1)(\boldsymbol{\beta}^*_{s'})^T\boldsymbol{\beta}^*_{s'} \\
\geq\ & -\sqrt{nb_2(M+1)M^2}\sqrt{8q\log(n^\eta p)} - \frac{1}{2}nb_2(M+1)M^2 \\
\geq\ & -nb_2(M+1)M^2
\end{aligned}$$

The last inequality is valid since for $n$ sufficiently large, $\sqrt{nb_2(M+1)M^2}\sqrt{8q\log(n^\eta p)} = o(nb_2(M+1)M^2)$

Notice $l(D_n|\widehat{\boldsymbol{\beta}}_{s,s'}) > l(D_n|\mathbf{0}_{s'})$, so we can combine the above two inequalities

together and obtain

$$\begin{aligned} -nb_2(M+1)M^2 &\leq l(D_n|\mathbf{0}_{s'}) - l(D_n|\boldsymbol{\beta}^*_{s'}) \leq l(D_n|\widehat{\boldsymbol{\beta}}_{s,s'}) - l(D_n|\boldsymbol{\beta}^*_{s'}) \\ &\leq -\frac{b_1(M+1)n}{4}||\widehat{\boldsymbol{\beta}}_{s,s'} - \boldsymbol{\beta}^*_{s'}||_2 \end{aligned}$$

which leads to

$$||\widehat{\boldsymbol{\beta}}_{s,s'} - \boldsymbol{\beta}^*_{s'}||_2 \leq \frac{4b_2(M+1)M^2}{b_1(M+1)}$$

and

$$||\widehat{\boldsymbol{\beta}}_{s,s'}||_2 \leq R$$

□

## C.2 Proof of Theorem 4.1

From Lemma C.1, we know for sufficiently large $n$, with probability at least $1 - n^{-\eta}$,

1. For all model s with $|s| \leq q$ and $s \not\supseteq s^*$, $||\widehat{\boldsymbol{\beta}}_s||_2 \leq R$ ,where $R = M + 1 + \frac{4b_2(M+1)M^2}{b_1(M+1)}$
2. For all model s with $|s| \leq q$ and $s \supseteq s^*$, $||\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}^*_s||_2 = o(1)$, therefore $||\widehat{\boldsymbol{\beta}}_s||_2 \leq R$ as well.

To sum up, for all model s with $|s| \leq q$, $||\widehat{\boldsymbol{\beta}}_s||_2$ is bounded by R.

Notice

$$
\begin{aligned}
E_{\{\boldsymbol{\beta}_s|D_n,s\}}\left[l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)\right] &= \int_{\boldsymbol{\beta}_s} P(\boldsymbol{\beta}_s|D_n,s)\left[l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)\right]d\boldsymbol{\beta}_s \\
&= \int_{\boldsymbol{\beta}_s} \frac{L(D_n|\boldsymbol{\beta}_s)\pi(\boldsymbol{\beta}_s)}{m(D_n|s)}\left[l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)\right]d\boldsymbol{\beta}_s \\
&= \frac{\int_{\boldsymbol{\beta}_s} L(D_n|\boldsymbol{\beta}_s)\pi(\boldsymbol{\beta}_s)\left[l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)\right]d\boldsymbol{\beta}_s}{m(D_n|s)}
\end{aligned}
$$

In the following, we deal with the numerator and denominator, respectively.

**Part 1: Numerator**

For the numerator, we split the integral domain into three regions, a small neighborhood of the MLE $\widehat{\boldsymbol{\beta}}_s$ denoted by $N_1$, the area between the small neighborhood $N_1$ and a larger neighborhood $N_2$, and the rest $\mathbb{R}^s \backslash N_2$. More specifically, $N_1$ and $N_2$ are defined as

$$
\begin{aligned}
N_1 &= \{\boldsymbol{\beta}_s : ||H(D_n, \widehat{\boldsymbol{\beta}}_s)^{1/2}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_s)||_2 \leq \sqrt{4\log(n)}\} \\
N_2 &= \{\boldsymbol{\beta}_s : ||H(D_n, \widehat{\boldsymbol{\beta}}_s)^{1/2}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_s)||_2 \leq \sqrt{nb_1(R+1)}\}
\end{aligned}
$$

Now the numerator can be written as a sum of three integrals.

$$
\begin{aligned}
& \int L(D_n|\boldsymbol{\beta}_s)\pi(\boldsymbol{\beta}_s)\left[l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)\right]d\boldsymbol{\beta}_s \\
= & \int_{\boldsymbol{\beta}_s \in N_1} L(D_n|\boldsymbol{\beta}_s)\pi(\boldsymbol{\beta}_s)\left[l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)\right]d\boldsymbol{\beta}_s \ (\textbf{Int1}) \\
+ & \int_{\boldsymbol{\beta}_s \in N_2 \backslash N_1} L(D_n|\boldsymbol{\beta}_s)\pi(\boldsymbol{\beta}_s)\left[l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)\right]d\boldsymbol{\beta}_s \ (\textbf{Int2}) \\
+ & \int_{\boldsymbol{\beta}_s \in \mathbb{R}^s \backslash N_2} L(D_n|\boldsymbol{\beta}_s)\pi(\boldsymbol{\beta}_s)\left[l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)\right]d\boldsymbol{\beta}_s \ (\textbf{Int3})
\end{aligned}
$$

**Int1**:

By applying part 1 of lemma C.1, for $\boldsymbol{\beta}_s \in N_1$, we have

$$\begin{aligned}\sqrt{4\log(n)} &\geq ||H(D_n, \widehat{\boldsymbol{\beta}}_s)^{1/2}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_s)||_2 \\ &= \sqrt{(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_s)^T H(D_n, \widehat{\boldsymbol{\beta}}_s)^{1/2}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_s)} \\ &\geq \sqrt{nb_1(R+1)}||\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_s||_2\end{aligned}$$

Thus $||\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_s||_2 \leq \sqrt{\frac{4\log(n)}{nb_1(R+1)}} < 1$

In this small neighborhood around $\widehat{\boldsymbol{\beta}}_s$, expand $l(D_n|\boldsymbol{\beta}_s)$ at $\widehat{\boldsymbol{\beta}}_s$, and we have

$$\begin{aligned}& l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s) \\ = \;& -\frac{1}{2}(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)^T\Big[H(D_n, \widehat{\boldsymbol{\beta}}_s + t(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s))\Big](\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s) \\ = \;& -\frac{1}{2}(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)^T\Big[H(D_n, \widehat{\boldsymbol{\beta}}_s)\Big](\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s) \\ & -\frac{1}{2}(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)^T\Big[H(D_n, \widehat{\boldsymbol{\beta}}_s + t(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)) - H(D_n, \widehat{\boldsymbol{\beta}}_s)\Big](\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)\end{aligned}$$

Since $||\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s||_2 < 1$, we have $||\widehat{\boldsymbol{\beta}}_s + t(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)||_2 \leq R + 1$. By applying part 1 in lemma C.1, we have

$$\begin{aligned}& \Big|(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)^T\Big\{H(D_n, \widehat{\boldsymbol{\beta}}_s + t(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)) - H(D_n, \widehat{\boldsymbol{\beta}}_s)\Big\}(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)\Big| \\ \leq \;& ntb_3(R+1)||\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s||_2^3 \leq nb_3(R+1)||\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s||_2^3\end{aligned}$$

and

$$\Big|(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)^T H(D_n, \widehat{\boldsymbol{\beta}}_s)(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)\Big| \geq nb_1(R+1)||\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s||_2^2$$

Combing them together, we further have

$$\begin{aligned} & \frac{\left|(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)^T \left\{ H(D_n, \widehat{\boldsymbol{\beta}}_s + t(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)) - H(D_n, \widehat{\boldsymbol{\beta}}_s) \right\} (\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s) \right|}{\left|(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)^T H(D_n, \widehat{\boldsymbol{\beta}}_s)(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)\right|} \\ \leq & \frac{b_3(R+1)}{b_1(R+1)} ||\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s||_2 \leq \sqrt{\frac{4\log(n) b_3^2(R+1)}{n b_1^3(R+1)}} \end{aligned}$$

In order to simplify the notation , we let $\xi_n = \sqrt{\frac{4\log(n) b_3^2(R+1)}{n b_1^3(R+1)}}$

Recall

$$\pi(\boldsymbol{\beta}_s) = N(0, V_s) = N(0, \sigma_s^2 I)$$

where $\sigma_s^2 = \frac{1}{2\pi} e^{C_0/|s|}$ for some positive constants $C_0$. It's easy to verify there exists three constants $K_1$, $K_2$ and $K_3$ such that for any model $s$ with $|s| \leq q$, we have

$$\begin{aligned} \sup_{\boldsymbol{\beta}_s} \pi(\boldsymbol{\beta}_s) \leq & \quad K_1 \quad < \infty \\ \inf_{||\boldsymbol{\beta}_s||_2 \leq R+1} \pi(\boldsymbol{\beta}_s) \geq & \quad K_2 \quad > 0 \\ \sup_{||\boldsymbol{\beta}_s||_2 \leq R+1} ||\nabla \pi(\boldsymbol{\beta}_s)||_2 \leq & \quad K_3 \quad < \infty \end{aligned}$$

These also imply

$$\sup_{||\boldsymbol{\beta}_s||_2 \leq R+1} ||\nabla \log \pi(\boldsymbol{\beta}_s)||_2 = \sup_{||\boldsymbol{\beta}_s||_2 \leq R+1} ||\frac{\nabla \pi(\boldsymbol{\beta}_s)}{\pi(\boldsymbol{\beta}_s)}||_2 \leq \frac{K_3}{K_2} < \infty$$

and

$$\sup_{||\boldsymbol{\beta}_s||_2 \leq R+1} |\log \pi(\boldsymbol{\beta}_s) - \log \pi(\widehat{\boldsymbol{\beta}}_s)| \leq \frac{K3}{K2} \sqrt{\frac{4\log(n)}{n b_1 (R+1)}}$$

Combining all the above analysis together, we can obtain the lower bound of Int1

$$
\begin{aligned}
& \int_{\boldsymbol{\beta}_s \in N_1} L(D_n|\boldsymbol{\beta}_s)\pi(\boldsymbol{\beta}_s)\Big[l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)\Big]d\boldsymbol{\beta}_s \\
= \;& \exp\{l(D_n|\widehat{\boldsymbol{\beta}}_s)\}\int_{\boldsymbol{\beta}_s \in N_1} \exp\{l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s) \\
& + \log\pi(\boldsymbol{\beta}_s)\}\Big[l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)\Big]d\boldsymbol{\beta}_s \\
\geq \;& \exp\{l(D_n|\widehat{\boldsymbol{\beta}}_s) + \log\pi(\widehat{\boldsymbol{\beta}}_s) + \frac{K3}{K2}\sqrt{\frac{4\log(n)}{nb_1(R+1)}}\} \\
& \times \int_{\boldsymbol{\beta}_s \in N_1} \exp\{-\frac{1}{2}(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)^T H(D_n, \widehat{\boldsymbol{\beta}}_s)(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)(1-\xi_n)\} \\
& \Big\{ -\frac{1}{2}(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)^T H(D_n, \widehat{\boldsymbol{\beta}}_s)(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)(1+\xi_n)\Big\}d\boldsymbol{\beta}_s
\end{aligned}
$$

Introduce new variable $\mathbf{t} = \sqrt{1-\xi_n}H(D_n, \widehat{\boldsymbol{\beta}}_s)^{1/2}(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)$, the lower bound becomes

$$
\begin{aligned}
& \exp\Big\{l(D_n|\widehat{\boldsymbol{\beta}}_s) + \log\pi(\widehat{\boldsymbol{\beta}}_s) + \frac{K3}{K2}\sqrt{\frac{4\log(n)}{nb_1(R+1)}}\Big\}(1-\xi_n)^{-|s|/2}|H(D_n, \widehat{\boldsymbol{\beta}}_s)|^{-1/2} \\
& \times \int_{||\mathbf{t}||_2 \leq \sqrt{1-\xi_n}\sqrt{4\log(n)}} \frac{1+\xi_n}{1-\xi_n}\exp\{-\frac{1}{2}\mathbf{t}^T\mathbf{t}\}\Big\{ -\frac{1}{2}\mathbf{t}^T\mathbf{t}\Big\}d\mathbf{t} \\
\geq \;& \exp\Big\{l(D_n|\widehat{\boldsymbol{\beta}}_s) + \log\pi(\widehat{\boldsymbol{\beta}}_s) + \frac{K3}{K2}\sqrt{\frac{4\log(n)}{nb_1(R+1)}} + \log(\frac{1+\xi_n}{1-\xi_n})\Big\} \\
& (1-\xi_n)^{-|s|/2}|H(D_n, \widehat{\boldsymbol{\beta}}_s)|^{-1/2}(2\pi)^{|s|/2}\int_{\mathbf{t}\in\mathbb{R}^{|s|}}(2\pi)^{-|s|/2}\exp\{-\frac{1}{2}\mathbf{t}^T\mathbf{t}\}\Big\{ -\frac{1}{2}\mathbf{t}^T\mathbf{t}\Big\}d\mathbf{t} \\
= \;& \exp\Big\{l(D_n|\widehat{\boldsymbol{\beta}}_s) + \log\pi(\widehat{\boldsymbol{\beta}}_s) + \frac{K3}{K2}\sqrt{\frac{4\log(n)}{nb_1(R+1)}} + \log(\frac{1+\xi_n}{1-\xi_n})\Big\} \\
& (1-\xi_n)^{-|s|/2}|H(D_n, \widehat{\boldsymbol{\beta}}_s)|^{-1/2}(2\pi)^{|s|/2}(-\frac{|s|}{2})
\end{aligned}
$$

The last equality is derived from the fact

$$\int_{\mathbf{t}\in\mathbb{R}^{|s|}} (2\pi)^{-|s|/2}\exp\{-\frac{1}{2}\mathbf{t}^T\mathbf{t}\}\Big\{\mathbf{t}^T\mathbf{t}\Big\}d\mathbf{t} = E(\chi^2(|s|)) = |s|$$

We can also obtain the upper bound of Int1

$$\begin{aligned}
& \int_{\boldsymbol{\beta}_s\in N_1} L(D_n|\boldsymbol{\beta}_s)\pi(\boldsymbol{\beta}_s)\Big[l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)\Big]d\boldsymbol{\beta}_s \\
= & \exp\{l(D_n|\widehat{\boldsymbol{\beta}}_s)\}\int_{\boldsymbol{\beta}_s\in N_1}\exp\{l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s) \\
& + \log\pi(\boldsymbol{\beta}_s)\}\Big[l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)\Big]d\boldsymbol{\beta}_s \\
\leq & \exp\{l(D_n|\widehat{\boldsymbol{\beta}}_s) + \log\pi(\widehat{\boldsymbol{\beta}}_s) - \frac{K3}{K2}\sqrt{\frac{4\log(n)}{nb_1(R+1)}}\} \\
& \times\int_{\boldsymbol{\beta}_s\in N_1}\exp\{-\frac{1}{2}(\boldsymbol{\beta}_s-\widehat{\boldsymbol{\beta}}_s)^T H(D_n,\widehat{\boldsymbol{\beta}}_s)(\boldsymbol{\beta}_s-\widehat{\boldsymbol{\beta}}_s)(1+\xi_n)\} \\
& \Big\{-\frac{1}{2}(\boldsymbol{\beta}_s-\widehat{\boldsymbol{\beta}}_s)^T H(D_n,\widehat{\boldsymbol{\beta}}_s)(\boldsymbol{\beta}_s-\widehat{\boldsymbol{\beta}}_s)(1-\xi_n)\Big\}d\boldsymbol{\beta}_s
\end{aligned}$$

Introduce new variable $\mathbf{t} = \sqrt{1+\xi_n}H(D_n,\widehat{\boldsymbol{\beta}}_s)^{1/2}(\boldsymbol{\beta}_s-\widehat{\boldsymbol{\beta}}_s)$, the upper bound becomes

$$\begin{aligned}
& \exp\Big\{l(D_n|\widehat{\boldsymbol{\beta}}_s) + \log\pi(\widehat{\boldsymbol{\beta}}_s) - \frac{K3}{K2}\sqrt{\frac{4\log(n)}{nb_1(R+1)}}\Big\}(1+\xi_n)^{-|s|/2}|H(D_n,\widehat{\boldsymbol{\beta}}_s)|^{-1/2} \\
& \times\int_{||\mathbf{t}||_2\leq\sqrt{1+\xi_n}\sqrt{4\log(n)}}\frac{1-\xi_n}{1+\xi_n}\exp\{-\frac{1}{2}\mathbf{t}^T\mathbf{t}\}\Big\{-\frac{1}{2}\mathbf{t}^T\mathbf{t}\Big\}d\mathbf{t} \\
\leq & \exp\Big\{l(D_n|\widehat{\boldsymbol{\beta}}_s) + \log\pi(\widehat{\boldsymbol{\beta}}_s) - \frac{K3}{K2}\sqrt{\frac{4\log(n)}{nb_1(R+1)}} + \log(\frac{1-\xi_n}{1+\xi_n})\Big\} \\
& (1+\xi_n)^{-|s|/2}|H(D_n,\widehat{\boldsymbol{\beta}}_s)|^{-1/2}(2\pi)^{|s|/2}\int_{||\mathbf{t}||_2\leq\sqrt{4\log(n)}}(2\pi)^{-|s|/2}\exp\{-\frac{1}{2}\mathbf{t}^T\mathbf{t}\}d\mathbf{t} \\
\leq & \exp\Big\{l(D_n|\widehat{\boldsymbol{\beta}}_s) + \log\pi(\widehat{\boldsymbol{\beta}}_s) - \frac{K3}{K2}\sqrt{\frac{4\log(n)}{nb_1(R+1)}} + \log(\frac{1-\xi_n}{1+\xi_n})\Big\} \\
& (1+\xi_n)^{-|s|/2}|H(D_n,\widehat{\boldsymbol{\beta}}_s)|^{-1/2}(2\pi)^{|s|/2}(-\frac{|s|}{2})(1-\frac{2^{|s|/2}n^{-1/2}}{|s|/2})
\end{aligned}$$

The last inequality is based on the observation that for $n$ sufficiently large and $||\mathbf{t}||_2 \geq \sqrt{4\log(n)}$, we have $\frac{1}{2}\mathbf{t}^T\mathbf{t} \leq \exp\{\frac{1}{4}\mathbf{t}^T\mathbf{t}\}$, so

$$\begin{aligned}
& \int_{||\mathbf{t}||_2 \geq \sqrt{4\log(n)}} (2\pi)^{-|s|/2} \exp\{-\frac{1}{2}\mathbf{t}^T\mathbf{t}\} \left\{ \frac{1}{2}\mathbf{t}^T\mathbf{t} \right\} d\mathbf{t} \\
\leq & \int_{||\mathbf{t}||_2 \geq \sqrt{4\log(n)}} (2\pi)^{-|s|/2} \exp\{-\frac{1}{4}\mathbf{t}^T\mathbf{t}\} d\mathbf{t} \\
= & \ 2^{|s|/2} \int_{||\mathbf{t}||_2 \geq \sqrt{4\log(n)}} (2\pi)^{-|s|/2} (2)^{-|s|/2} \exp\{-\frac{1}{2}\mathbf{t}^T(2I)^{-1}\mathbf{t}\} \\
= & \ 2^{|s|/2} P(\chi^2(|s|) \geq 2\log(n)) \leq 2^{|s|/2} e^{-\log(n)/2} = 2^{|s|/2} n^{-1/2}
\end{aligned}$$

By carefully reorganizing the terms in upper bound and lower bound, we can finally obtain that for $n$ sufficiently large, there exists two positive constants $c_1$ and $c_2$ such that

$$\begin{aligned}
\text{Int1} &\leq -\frac{|s|}{2} L(D_n, \widehat{\boldsymbol{\beta}}_s) \pi(\widehat{\boldsymbol{\beta}}_s) |H(D_n, \widehat{\boldsymbol{\beta}}_s)|^{-1/2} (2\pi)^{|s|/2} \left\{ 1 + c_1 \sqrt{\frac{\log(n)}{n}} \right\} \\
\text{Int1} &\geq -\frac{|s|}{2} L(D_n, \widehat{\boldsymbol{\beta}}_s) \pi(\widehat{\boldsymbol{\beta}}_s) |H(D_n, \widehat{\boldsymbol{\beta}}_s)|^{-1/2} (2\pi)^{|s|/2} \left\{ 1 - c_2 \sqrt{\frac{\log(n)}{n}} \right\}
\end{aligned}$$

**Int2**

By applying part 1 of lemma C.1, for $\boldsymbol{\beta}_s \in N_2$, we have

$$\begin{aligned}
\sqrt{b_1(R+1)n} &\geq ||H(D_n, \widehat{\boldsymbol{\beta}}_s)^{1/2} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_s)||_2 \\
&= \sqrt{(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_s)^T H(D_n, \widehat{\boldsymbol{\beta}}_s)^{1/2} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_s)} \\
&\geq \sqrt{n b_1 (R+1)} ||\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_s||_2
\end{aligned}$$

Thus $||\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s||_2 \leq 1$. Apply part 1 of lemma C.1 again and we obtain

$$\begin{aligned} |l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)| &= |-\frac{1}{2}(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)^T H(D_n, \widehat{\boldsymbol{\beta}}_s + t(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s))(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)| \\ &\geq \frac{1}{2}\frac{b_1(R+1)}{b_2(R+1)}(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)^T H(D_n, \widehat{\boldsymbol{\beta}}_s)(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s) \end{aligned}$$

It is also easy to verify that for $n$ sufficiently large,

$$|l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)| \leq \exp\{\frac{1}{2}|l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)|\}$$

Based on these information, now let's deal with Int2

$$\begin{aligned} & |\text{Int2}| \\ = & \left| \int_{\boldsymbol{\beta}_s \in N_2} L(D_n|\boldsymbol{\beta}_s)\pi(\boldsymbol{\beta}_s)[l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)]d\boldsymbol{\beta}_s \right| \\ = & \exp\{l(D_n|\widehat{\boldsymbol{\beta}}_s)\} \int_{\boldsymbol{\beta}_s \in N_2} \exp\{l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)\}\pi(\boldsymbol{\beta}_s)[l(D_n|\widehat{\boldsymbol{\beta}}_s) - l(D_n|\boldsymbol{\beta}_s)]d\boldsymbol{\beta}_s \\ \leq & \exp\{l(D_n|\widehat{\boldsymbol{\beta}}_s)\} \int_{\boldsymbol{\beta}_s \in N_2} \exp\{\frac{1}{2}[l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)]\}\pi(\boldsymbol{\beta}_s)d\boldsymbol{\beta}_s \\ \leq & K_1 \exp\{l(D_n|\widehat{\boldsymbol{\beta}}_s)\} \int_{\boldsymbol{\beta}_s \in N_2} \exp\{-\frac{1}{4}\frac{b_1(R+1)}{b_2(R+1)}(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)^T H(D_n, \widehat{\boldsymbol{\beta}}_s)(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)\}d\boldsymbol{\beta}_s \\ \leq & K_1 \exp\{l(D_n|\widehat{\boldsymbol{\beta}}_s)\} \times \\ & \int_{||(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_s)||_2 \geq \sqrt{4\log(n)}} \exp\{-\frac{1}{4}\frac{b_1(R+1)}{b_2(R+1)}(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)^T H(D_n, \widehat{\boldsymbol{\beta}}_s)(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)\}d\boldsymbol{\beta}_s \end{aligned}$$

Let $\mathbf{t} = \sqrt{\frac{b_1(R+1)}{2b_2(R+1)}} H(D_n, \widehat{\boldsymbol{\beta}}_s)^{1/2}(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)$, then the upper bound becomes

$$\begin{aligned}
& K_1 \exp\{l(D_n|\widehat{\boldsymbol{\beta}}_s)\}(\frac{b_1(R+1)}{2b_2(R+1)})^{-|s|/2}|H(D_n, \widehat{\boldsymbol{\beta}}_s|^{-1/2}(2\pi)^{|s|/2} \\
& \times \int_{||\mathbf{t}||_2 \geq \sqrt{\frac{2b_1(R+1)\log(n)}{b_2(R+1)}}} (2\pi)^{-|s|/2} \exp\{-\frac{1}{2}\mathbf{t}^T\mathbf{t}\} d\mathbf{t} \\
= \quad & K_1 \exp\{l(D_n|\widehat{\boldsymbol{\beta}}_s)\}(\frac{b_1(R+1)}{2b_2(R+1)})^{-|s|/2}|H(D_n, \widehat{\boldsymbol{\beta}}_s|^{-1/2}(2\pi)^{|s|/2} \\
& P(\chi^2(|s|) \geq \frac{2b_1(R+1)\log(n)}{b_2(R+1)}) \\
\leq \quad & K_1 \exp\{l(D_n|\widehat{\boldsymbol{\beta}}_s)\}(\frac{b_1(R+1)}{2b_2(R+1)})^{-|s|/2}|H(D_n, \widehat{\boldsymbol{\beta}}_s|^{-1/2}(2\pi)^{|s|/2} n^{-\frac{b_1(R+1)}{2b_2(R+1)}}
\end{aligned}$$

By carefully reorganizing the terms above, we can finally obtain that for $n$ sufficiently large, there exists a positive constant $c_3$ such that

$$|\text{Int2}| \leq \frac{|s|}{2} L(D_n, \widehat{\boldsymbol{\beta}}_s)\pi(\widehat{\boldsymbol{\beta}}_s)|H(D_n, \widehat{\boldsymbol{\beta}}_s)|^{-1/2}(2\pi)^{|s|/2}\left\{c_3 n^{-\frac{b_1(R+1)}{2b_2(R+1)}}\right\}$$

**Int3**

For $||H^{1/2}(D_n, \widehat{\boldsymbol{\beta}}_s)(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)||_2 = \sqrt{b_1(R+1)n}$, we have $||\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s||_2 \leq 1$, thus

$$\begin{aligned}
|l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)| \quad & \geq \quad \frac{1}{2}\frac{b_1(R+1)}{b_2(R+1)}(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)^T H(D_n, \widehat{\boldsymbol{\beta}}_s)(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s) \\
& = \quad \frac{b_1^{3/2}(R+1)\sqrt{n}}{2b_2(R+1)}||H^{1/2}(D_n, \widehat{\boldsymbol{\beta}}_s)(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)||_2
\end{aligned}$$

By concavity of log-likelihood, for $||H^{1/2}(D_n, \widehat{\boldsymbol{\beta}}_s)(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)||_2 \geq \sqrt{b_1(R+1)n}$, we also have

$$|l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)| \geq \frac{b_1^{3/2}(R+1)\sqrt{n}}{2b_2(R+1)}||H^{1/2}(D_n, \widehat{\boldsymbol{\beta}}_s)(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)||_2$$

Another important observation is that for $n$ sufficiently large, for $||H^{1/2}(D_n, \widehat{\boldsymbol{\beta}}_s)(\boldsymbol{\beta}_s -$

$\widehat{\boldsymbol{\beta}}_s)||_2 \geq \sqrt{b_1(R+1)n}$, the following holds

$$|l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)| \leq \exp\{\frac{1}{2}|l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)|\}$$

Consequently,

$$\begin{aligned}
& |\text{Int3}| \\
= & \left| \int_{\boldsymbol{\beta}_s \in N_3} L(D_n|\boldsymbol{\beta}_s)\pi(\boldsymbol{\beta}_s)[l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)]d\boldsymbol{\beta}_s \right| \\
= & \exp\{l(D_n|\widehat{\boldsymbol{\beta}}_s)\} \int_{\boldsymbol{\beta}_s \in N_3} \exp\{l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)\}\pi(\boldsymbol{\beta}_s)[l(D_n|\widehat{\boldsymbol{\beta}}_s) - l(D_n|\boldsymbol{\beta}_s)]d\boldsymbol{\beta}_s \\
\leq & \exp\{l(D_n|\widehat{\boldsymbol{\beta}}_s)\} \int_{\boldsymbol{\beta}_s \in N_3} \exp\{\frac{1}{2}[l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)]\}\pi(\boldsymbol{\beta}_s)d\boldsymbol{\beta}_s \\
\leq & K_1 \exp\{l(D_n|\widehat{\boldsymbol{\beta}}_s)\} \times \\
& \int_{||(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_s)||_2 \geq \sqrt{b_1(R+1)n}} \exp\{-\frac{b_1^{3/2}(R+1)\sqrt{n}}{4b_2(R+1)}||H^{1/2}(D_n, \widehat{\boldsymbol{\beta}}_s)(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)||_2\}d\boldsymbol{\beta}_s
\end{aligned}$$

Let $\xi_n = \frac{b_1^{3/2}(R+1)\sqrt{n}}{4b_2(R+1)}$ and $\mathbf{t} = \xi_n H^{1/2}(D_n, \widehat{\boldsymbol{\beta}}_s)(\boldsymbol{\beta}_s - \widehat{\boldsymbol{\beta}}_s)$ , now the upper bound becomes

$$K_1 \exp\{l(D_n|\widehat{\boldsymbol{\beta}}_s)\}|H(D_n, \widehat{\boldsymbol{\beta}}_s)|^{-1/2}(\xi_n)^{-1} \int_{||\mathbf{t}||_2 \geq \xi_n\sqrt{b_1(R+1)n}} \exp\{-||\mathbf{t}||_2\}d\mathbf{t}$$

From Lemma2 in Foygel Barber *et al.* (2015), for $n$ sufficiently large, we have

$$\begin{aligned}
& \int_{||\mathbf{t}||_2 \geq \xi_n\sqrt{b_1(R+1)n}} \exp\{-||\mathbf{t}||_2\}d\mathbf{t} \\
\leq & \frac{4(\pi)^{|s|/2}[\xi_n\sqrt{b_1(R+1)n}]^{|s|-1}}{\Gamma(|s|/2)} \exp\{-\xi_n\sqrt{b_1(R+1)n}\} \\
\leq & \exp\{-\frac{\xi_n}{2}\sqrt{b_1(R+1)n}\}
\end{aligned}$$

By carefully reorganizing the terms above, we can obtain that for $n$ sufficiently

large, there exists two positive constants $c_4$ and $c_5$ such that

$$|\text{Int3}| \leq \frac{|s|}{2} L(D_n, \widehat{\boldsymbol{\beta}}_s)\pi(\widehat{\boldsymbol{\beta}}_s)|H(D_n, \widehat{\boldsymbol{\beta}}_s)|^{-1/2}(2\pi)^{|s|/2}\Big\{c_4 \exp\{-c_5 n\}\Big\}$$

After analyzing the above three integrals one by one, we can finally add them together and have

$$\begin{aligned} & \int L(D_n|\boldsymbol{\beta}_s)\pi(\boldsymbol{\beta}_s)\Big[l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)\Big]d\boldsymbol{\beta}_s \\ = & -\frac{|s|}{2} L(D_n, \widehat{\boldsymbol{\beta}}_s)\pi(\widehat{\boldsymbol{\beta}}_s)|H(D_n, \widehat{\boldsymbol{\beta}}_s)|^{-1/2}(2\pi)^{|s|/2}\Big\{1 + o(1)\Big\} \end{aligned}$$

**Part 2: Denominator**

The denominator can be processed similarly to the numerator. Here we omit the details and only list the result:

$$m(D_n|s) = \int L(D_n|\boldsymbol{\beta}_s)\pi(\boldsymbol{\beta}_s)d\boldsymbol{\beta}_s = L(D_n, \widehat{\boldsymbol{\beta}}_s)\pi(\widehat{\boldsymbol{\beta}}_s)|H(D_n, \widehat{\boldsymbol{\beta}}_s)|^{-1/2}(2\pi)^{|s|/2}\Big\{1 + o(1)\Big\}$$

**Part 3: Combination Step**

At last, we have

$$\begin{aligned} & E_{\{\boldsymbol{\beta}_s|D_n,s\}}\Big[l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)\Big] \\ = & \frac{\int_{\boldsymbol{\beta}_s} L(D_n|\boldsymbol{\beta}_s)\pi(\boldsymbol{\beta}_s)\Big[l(D_n|\boldsymbol{\beta}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_s)\Big]d\boldsymbol{\beta}_s}{m(D_n|s)} \\ = & \frac{-\frac{|s|}{2} L(D_n, \widehat{\boldsymbol{\beta}}_s)\pi(\widehat{\boldsymbol{\beta}}_s)|H(D_n, \widehat{\boldsymbol{\beta}}_s)|^{-1/2}(2\pi)^{|s|/2}\Big\{1 + o(1)\Big\}}{L(D_n, \widehat{\boldsymbol{\beta}}_s)\pi(\widehat{\boldsymbol{\beta}}_s)|H(D_n, \widehat{\boldsymbol{\beta}}_s)|^{-1/2}(2\pi)^{|s|/2}\Big\{1 + o(1)\Big\}} \\ = & -\frac{|s|}{2}[1 + o(1)] = -\frac{|s|}{2} + o(1) \end{aligned}$$

which ends the proof.

### C.3 Proof of Theorem 4.2

Proof of theorem 4.2 is based on the proof of theorem 2 in Foygel and Drton (2011). Define $A_0 = \{s : s^* \subset s, |s| \leq q\}$ and $A_1 = \{s : s^* \nsubseteq s, |s| \leq q\}$, then we consider two cases

**Case 1: $\mathbf{s \in A_0}$**

For $n$ sufficiently large, apply lemma C.1 and we have

$$\begin{aligned}
& l(D_n|\widehat{\boldsymbol{\beta}}_s) - l(D_n|\boldsymbol{\beta}_s^*) \\
= \; & (\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_s^*)^T \mathbf{s}(D_n, \boldsymbol{\beta}_s^*) - \frac{1}{2}(\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_s^*)^T H(D_n, \boldsymbol{\beta}_s^* + t(\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_s^*))(\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_s^*) \\
\leq \; & (\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_s^*)^T \mathbf{s}(D_n, \boldsymbol{\beta}_s^*) - \frac{1}{2}(\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_s^*)^T H(D_n, \boldsymbol{\beta}_s^*)(\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_s^*) \\
& -\frac{1}{2}(\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_s^*)^T \Big\{ H(D_n, \boldsymbol{\beta}_s^* + t(\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_s^*)) - H(D_n, \boldsymbol{\beta}_s^*) \Big\} (\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_s^*) \\
\leq \; & (\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_s^*)^T \mathbf{s}(D_n, \boldsymbol{\beta}_s^*) - \frac{1}{2}(\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_s^*)^T H(D_n, \boldsymbol{\beta}_s^*)(\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_s^*) \\
& +\frac{1}{2} n b_3 (R+1) ||\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_s^*||_2^3 \\
\leq \; & \sup_{\boldsymbol{\beta}_s} \Big\{ (\boldsymbol{\beta}_s - \boldsymbol{\beta}_s^*)^T \mathbf{s}(D_n, \boldsymbol{\beta}_s^*) - \frac{1}{2}(\boldsymbol{\beta}_s - \boldsymbol{\beta}_s^*)^T H(D_n, \boldsymbol{\beta}_s^*)(\boldsymbol{\beta}_s - \boldsymbol{\beta}_s^*) \Big\} \\
& +\frac{1}{2} n b_3 (R+1) \Big\{ \tau \sqrt{\frac{\log n^\eta p}{n}} \Big\}^3
\end{aligned}$$

Notice $\sup_{\boldsymbol{\beta}_s} \Big\{ (\boldsymbol{\beta}_s - \boldsymbol{\beta}_s^*)^T \mathbf{s}(D_n, \boldsymbol{\beta}_s^*) - \frac{1}{2}(\boldsymbol{\beta}_s - \boldsymbol{\beta}_s^*)^T H(D_n, \boldsymbol{\beta}_s^*)(\boldsymbol{\beta}_s - \boldsymbol{\beta}_s^*) \Big\}$ is achieved when $\boldsymbol{\beta}_s - \boldsymbol{\beta}_s^* = H(D_n, \boldsymbol{\beta}_s^*)^{-1} \mathbf{s}(D_n, \boldsymbol{\beta}_s^*)$. Plug it into the expression, we can obtain

$$\begin{aligned}
& \sup_{\boldsymbol{\beta}_s} \Big\{ (\boldsymbol{\beta}_s - \boldsymbol{\beta}_s^*)^T \mathbf{s}(D_n, \boldsymbol{\beta}_s^*) - \frac{1}{2}(\boldsymbol{\beta}_s - \boldsymbol{\beta}_s^*)^T H(D_n, \boldsymbol{\beta}_s^*)(\boldsymbol{\beta}_s - \boldsymbol{\beta}_s^*) \Big\} \\
= \; & \frac{1}{2} \mathbf{s}(D_n, \boldsymbol{\beta}_s^*) H(D_n, \boldsymbol{\beta}_s^*)^{-1} \mathbf{s}(D_n, \boldsymbol{\beta}_s^*)
\end{aligned}$$

Then

$$
\begin{aligned}
& l(D_n|\widehat{\boldsymbol{\beta}}_s) - l(D_n|\boldsymbol{\beta}^*_s) \\
\leq\ & \frac{1}{2}\mathbf{s}(D_n,\boldsymbol{\beta}^*_s)H(D_n,\boldsymbol{\beta}^*_s)^{-1}\mathbf{s}(D_n,\boldsymbol{\beta}^*_s) + \frac{1}{2}nb_3(R+1)\left\{\tau\sqrt{\frac{\log n^\eta p}{n}}\right\}^3 \\
\leq\ & (1+\epsilon_n)|s\backslash s^*|\log(n^\eta p) + \frac{b_3(R+1)\tau^3\sqrt{\log n^\eta p}}{2\sqrt{n}}\log(n^\eta p) \\
=\ & (1+o(1))|s\backslash s^*|\log(n^\eta p)
\end{aligned}
$$

Therefore, for $n$ sufficiently large

$$
\begin{aligned}
& AEBIC(s) - AEBIC(s^*) \\
=\ & -2E_{\{\boldsymbol{\beta}_s|D_n,s\}}l(D_n|\boldsymbol{\beta}_s) + 2E_{\{\boldsymbol{\beta}_{s^*}|D_n,s\}}l(D_n|\boldsymbol{\beta}_{s^*}) + |s\backslash s^*|\log(n) + 2\gamma|s\backslash s^*|\log(p) \\
\geq\ & -2\left[l(D_n|\widehat{\boldsymbol{\beta}}_s) - \frac{|s|}{2}+1\right] + 2\left[l(D_n|\widehat{\boldsymbol{\beta}}_{s^*}) - \frac{|s^*|}{2}-1\right] \\
& +|s\backslash s^*|\log(n) + 2\gamma|s\backslash s^*|\log(p) \\
\geq\ & -2\left[l(D_n|\widehat{\boldsymbol{\beta}}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_{s^*})\right] + |s\backslash s^*|\log(n) + 2\gamma|s\backslash s^*|\log(p) - 4 \\
\geq\ & -2\left[l(D_n|\widehat{\boldsymbol{\beta}}_s) - l(D_n|\boldsymbol{\beta}^*_s)\right] + |s\backslash s^*|\log(n) + 2\gamma|s\backslash s^*|\log(p) - 4 \\
\geq\ & -2(1+o(1))|s\backslash s^*|\log(n^\eta p) + |s\backslash s^*|\log(n) + 2\gamma|s\backslash s^*|\log(p) - 4 \\
=\ & -2|s\backslash s^*|\left\{(1+o(1))\log(n^\eta p) - \log(n^{1/2}p^\gamma)\right\} - 4
\end{aligned}
$$

Recall $p = o(n^\kappa)$ and $\eta < \frac{1}{2}$

$$
\begin{aligned}
\log(n^\eta p) - \log(n^{1/2}p^\gamma) &= (\eta - \frac{1}{2})\log(n) + (1-\gamma)\log(p) \\
&\leq (\eta - \frac{1}{2})\frac{1}{\kappa}\log(p) + (1-\gamma)\log(n) \\
&= (\frac{\eta}{\kappa} - \frac{1}{2\kappa} + 1 - \gamma)\log(n)
\end{aligned}
$$

Note also $\gamma > 1 - \frac{1-2\eta}{2\kappa}$ implies $\frac{\eta}{\kappa} - \frac{1}{2\kappa} + 1 - \gamma < 0$,. Therefore we finally proved

$$AEBIC(s) - AEBIC(s^*) > 0$$

for $n$ sufficiently large

**Case 2: $\mathbf{s \in A_1}$**

Let $s' = s \cup s^*$ and $\hat{\boldsymbol{\beta}}_{s,s'}$ denote a vector corresponding to model $s'$, generated by $\hat{\boldsymbol{\beta}}_s$ augmented with zeros in $s' \backslash s$. By applying part 3 of lemma C.1, we have

$$\begin{aligned} & l(D_n, \widehat{\boldsymbol{\beta}}_{s,s'}) - l(D_n, \boldsymbol{\beta}^*_{s'}) \\ \leq \; & -\frac{b_1(M+1)n}{2} ||\widehat{\boldsymbol{\beta}}_{s,s'} - \boldsymbol{\beta}^*_{s'}||_2 \left[ \min(1, ||\widehat{\boldsymbol{\beta}}_{s,s'} - \boldsymbol{\beta}^*_{s'}||_2) - \tau \sqrt{\frac{\log(n^\eta p)}{n}} \right] \end{aligned}$$

Since $\min_{j \in s^*} |\beta^*_j| \geq c_0 n^{-1/4}$. For $n$ sufficiently large, we have

$$\begin{aligned} & l(D_n, \widehat{\boldsymbol{\beta}}_{s,s'}) - l(D_n, \boldsymbol{\beta}^*_{s'}) \\ \leq \; & -\frac{b_1(M+1)n}{2} \min_{j \in s^*} |\beta^*_j| \left[ \min(1, \min_{j \in s^*} |\beta^*_j|) - \tau \sqrt{\frac{\log(n^\eta p)}{n}} \right] \\ \leq \; & -\frac{b_1(M+1)n}{2} c_0 (n^{-1/4}) (\frac{1}{2} c_0 n^{-1/4}) = -\frac{b_1(M+1)c_0^2 \sqrt{n}}{4} \end{aligned}$$

Therefore, for $n$ sufficiently large

$$
\begin{aligned}
& AEBIC(s) - AEBIC(s^*) \\
= \; & -2E_{\{\boldsymbol{\beta}_s|D_n,s\}} l(D_n|\boldsymbol{\beta}_s) + 2E_{\{\boldsymbol{\beta}_{s^*}|D_n,s\}} l(D_n|\boldsymbol{\beta}_{s^*}) \\
& +(|s| - |s^*|)\log(n) + 2(|s| - |s^*|)\gamma\log(p) \\
\geq \; & -2\left[l(D_n|\widehat{\boldsymbol{\beta}}_s) - \frac{|s|}{2} + 1\right] + 2\left[l(D_n|\widehat{\boldsymbol{\beta}}_{s^*}) - \frac{|s^*|}{2} - 1\right] \\
& +(|s| - |s^*|)\log(n) + 2(|s| - |s^*|)\gamma\log(p) \\
\geq \; & -2\left[l(D_n|\widehat{\boldsymbol{\beta}}_s) - l(D_n|\widehat{\boldsymbol{\beta}}_{s^*})\right] + (|s| - |s^*|)\log(n) + 2(|s| - |s^*|)\gamma\log(p) - 4 - q \\
\geq \; & -2\left[l(D_n|\widehat{\boldsymbol{\beta}}_s) - l(D_n|\boldsymbol{\beta}_s^*)\right] + (|s| - |s^*|)\log(n) + 2(|s| - |s^*|)\gamma\log(p) - 4 - q \\
\geq \; & \frac{b_1(M+1)c_0^2\sqrt{n}}{2} - q\log(n) - 2q\gamma\log(p) - 4 - q
\end{aligned}
$$

Simple analysis reveals

$$AEBIC(s) - AEBIC(s^*) > 0$$

for $n$ sufficiently large